HARVEY

378.744
BO
AM1932
harv

Boston University
College of Liberal Arts
Library

THE GIFT OF the Author

378.744
BO
AM1932

48504 harv

B O S T O N U N I V E R S I T Y

G R A D U A T E S C H O O L

Thesis

TEACHING HISTORY, THROUGH TESTING, IN THE HIGH SCHOOL

by

Carl Upton Harvey

(B.S. in Ed. Boston University, 1931)

submitted in partial fulfilment of the

requirements for the degree of

Master of Arts

1932

BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

48504

378.744
BO
A.M.1932
harv

CONTENTS

Teaching History, through Testing, in the High School

Chapter 1

Introductory

1) Need for the study of measuring knowledge and achievement in history---------- 2

2) The principal types of measurements in history in the modern high school---------- 3

Chapter 2

Uses of History Tests in the High School

1) History tests may be used for examinations---- 6

2) History tests may measure a teacher's efficiency-7

3) Use of history tests to give pupils objective standards---------- 7

4) Diagnosis is a worthy use of history tests----- 8

5) History tests help in placement and classification---------- ---------- 9

6) History tests may be used for review and recall- 10

7) History tests enhance the intrinsic worth of learning---------- 12

8) History tests may measure memory---------- 14

9) History tests may be used for comparison----- 15

10) Use of history tests to improve teaching------ 15

11) Summary---------- 17

CONTENTS

Chapter 3

Criteria for the Selection of History Tests in the High School

1) Criteria to be considered---------- 20

2) Validity---------- 21

3) Reliability---------- 26

4) Other things being equal, choose a test that has norms---------- 32

5) Other things being equal, choose a test having alternative forms---------- 33

6) Choose a test because it is scaled---------- 35

7) Ease of administration---------- 37

8) Other things being equal, choose a test that is inexpensive---------- 39

9) Summary---------- 41

Chapter 4

Teaching History through the Use of Oral Questioning

1) The uses of oral questioning---------- 44

2) Oral versus written tests---------- 45

3) Conclusion---------- 47

Chapter 5

Teaching History through the Use of the Traditional Examination

1) Nature of the traditional examination---------- 49

CONTENTS

Chapter 5 (Continued)

2) Advantages of the traditional examination----- 51

3) Dis-advantages of the traditional examination---54

4) Conclusion----------67

Chapter 6

Teaching History through the Use of the Standard Test

1) Growth of the standard test---------- 70

2) Nature and purpose of the standard test---------71

3) Additional advantages of the standard test------76

4) Some limitations of the standard test----------77

5) Example of the standard test----------79

6) Conclusion----------82

Chapter 7

Teaching History through the Use of the Objective or New-Type Examination

1) Nature of the new-type examination---------- 85

2) Advantages of the new-type examination------- 86

3) Dis-advantages of the new-type examination---- 100

4) Five most generally useful of the new-type examinations ~~forms~~---------- 107

5) Conclusion---------- 110

CONTENTS

Chapter 8

Summary---------- 113

Appendix---------- 123

Teaching History, through Testing, in the High School.

Chapter 1

Introductory

1) Need for the study of measuring knowledge and achievement in history.
2) The Principal types of measurements in history in the modern high school.

Teaching History, through Testing, in the High School.

Chapter 1

Introductory

1) Need for the study of measuring knowledge and achievement in history-----

The value of educational measurement in general is no longer a matter of doubt. The use of tests and scales in thousands of schools has shown very clearly that from the results of testing there have come inestimable benefits to school administrators and teachers, to pupils in all grades and of all agesm and to the community at large. The school system in which educational measurement is widely used is likely to be markedly superior to the one in which there is no testing program.

The past quarter of a century has witnessed the rise of educational measurement to the plane of conscious striving for objective, impartial and comparative means for portraying the absolute and relative achievements of pupils. The measurement of achievement has been admittedly the principal reason for examinations. This idea is undoubtedly sound. It may require certain

qualification, but all seem to agree that the first purpose of a test or examination is that of ascertaining the degree to which individual pupils have profited by instruction.

It is plain that, although scientific measurement is still a very new thing in education, it has passed the experimental stage. It has become so important that no teacher can be said to have had complete professional training without some knowledge of this subject. Thus the teacher who is thoroughly equipped for her position must know something of the aims, methods, materials, results, and general educational implications of this important work.

2) The Principal types of measurements in history in the modern high school -----

Four types of measurements in history exist side by side in the modern high school. These are:

1. Oral questioning.
2. The traditional examination.
3. The standard test.
4. The objective or new-type examination.

There are other means of evaluating school results, but the four types mentioned are the most important. With such a variety of methods open to the educator, and with so little of a final character known about

their relative merits, the only course open to us is to consider both the logic and the growing body of experimental findings supporting or undermining the value of each. This is the task of this thesis in its entirety.

Chapter 2

Uses of History Tests in the High School

1) History tests may be used for examinations.
2) History tests may measure a teacher's efficiency.
3) Use of history tests to give pupils objective standards.
4) Diagnosis is a worthy use of history tests.
5) History tests help in placement and classification.
6) History tests may be used for review and recall.
7) History tests enhance the intrinsic worth of learning.
8) History tests may measure memory.
9) History tests may be used for comparison.
10) Use of history tests to improve teaching.
11) Summary.

Chapter 2

Uses of History Tests in the High School [1,2,3.]

1) History tests may be used for examinations----

One of the widest uses which all forms of testinghave had in the past, and one of their more important uses in the future, lies in the field of examinations. In this sense, tests are used to find out the status of an individual pupil at any given time from the p oint of view of his achievement. It may be that the primary purpose of examining is to determine the fitness of a pupil for promotion. As long as promotion to a higher grade depends largely on the academic fitness of the individual for such promotion, - just so long is this method of examining for promotion just and right. The great need, however, is to make sure that the examinations which are used for this purpose really measure what is wanted, and do measure real achievement.

Examinations to determine the status of pupils who have just entered a school are also much needed. This, again, is a form of determination of the status of a pupil, not primarily as a check after teaching has been done, but rather to find out where further teaching should begin.

In these forms of examinations, tests can be used

---

1 - From Symonds, Measurement in Secondary Education pp. 1-2

2 - From Russell, Classroom Tests. pp. 4-11

3- From Ruch and Stoddard, Tests and Measurements in High School Instruction pp.8-44

to advantage in securing not only useful results but results which are justified.

2) History tests may measure a teacher's efficiency

History tests may also be used as a measure of a teacher's efficiency. It is true we cannot say that a teacher has any definite percentage of teaching efficiency, as we can measure the development of horse power in a gasoline engine, and it would probably be undesirable if we could. We can, however, through the analysis of a series of tests, say that a teacher is doing better with his pupils than teachers of similar children elsewhere, or as well, or more poorly. This, then, may become for the individual teacher a potentfactor in helping him to appreciate where improvements in his teaching can be made,and whereeffort should be best placed. Without necessarily knowing the absolute amount of efficiency with which a teacher teaches, it becomes possible in this way to increase efficiency by preventing unnecessary repetitions of teaching, or by providing more clearly for teaching needs.

3) Use of history tests to give pupils objective standards.

A further use of tests is one which tends to give to pupils some objective standard by which to judge

the character or the quality of the work which they have done.

Pupils immersed in the details of school study frequently find it difficult to appreciate the objectives of that study. For that reason they have little, save their own interest or inclination, by which to judge the relative importance of varying phases of a subject as they arise. Tests may furnish such objectives, (or at least a high grade of substitute for such objectives), by helping students to see a relative value, and to appreciate the necessity for accurate and complete knowledge.

In another way the same object may be accomplished through the fact that the tests may provide a motive for the study. If the motive is not that of passing the subject, it is probable that it is worthy and capable of furnishing a worthy objective. Objective standards and objective goals are superior to non-objective standards and unknown goals.

4) Diagnosis is a worthy use of history tests---

In addition to these uses, however, one of the worthy uses to which testing can be put lies in diagnosis. This is a field as broad as that of the school itself, but one to which, until the present, there has been but little attention paid.

Diagnosis of the difficulties in the learning process, of the difficulties for individual pupils

in the varying types of subject matter, and in the various phases of the school curriculum,- all this is one of the widest and most important fields of endeavor for the teacher.

Diagnosis is a field of constant inquiry, a field of highly focused endeavor, and a field of rapid change. Here it is that the teacher has to do with the kaleidoscopic changes of tastes and perception, of likes and dislikes, of feelings and emotions. Here it is that he has to work with the individual differences of his pupils and with their inherent abilities. No adequate method by which these differences could be measured, no fair treatment of the pupils of a class, was possible, until tests and the method of measurement as we know it today were developed.

The diagnosis of individual difficulties and the remedial measures which must be taken to counteract these difficulties lie within the province of the teacher. They can be accomplished only through the medium of adequate testing.

5) History tests help in placement and classification

Another of the uses of history tests which is being rapidly developed at the present time is for the proper placement of new pupils in a school, and the reclassification of pupils already there.

On the other hand, pupils coming into a new school-district, or into a new school system, may have been taught by different standards, or possibly with a different curriculum, from the pupils in the schools which they are entering. History tests can be used to advantage with these prospective pupils to determine notonly their absolute achievement, but also their abilities in relation to the other pupils in the school. The results of these tests provides justification for making whatever placements are made.

On the other hand, many pupils, under the former basis of promotion, were frequently misplaced several grades although they were more often placed in grades below their actual school level than above it. In this case, a battery of tests administered throughout a school may be an objective determiner of the status of all pupils in the school, and it may also be a valid reason for changing the grade classification of many of them.

6) History tests may be used for review and recall-

One of the wider uses of testing is that of testing for purposes of review, or to help in the process of recall.

Psychologists have made many studies of the rate at which various subjects and skills are forgotten

after they have once been learned. In conformance with the "Laws of Learning" it has been found that, without exception, whatever is learned tends to deteriorate, other things being equal, with disuse, but that also each review or recall makes the learnedbonds stronger and stronger, until, with enough repetition of the right kind, the bonds may become so well established above the threshold of memory as to be relatively permanent.

Under such conditions it would seem to be one of the great aims of the teacher not only to teach, but also to reteach so as to increase the retention of the learned elements. Some types of teacher's tests help to make the subject matter so vivid and so full of interest that it may not only be retained but also be retained in a psychologically desirable way.

In addition to retention, however, there are also other purposes in a review which are important in teaching, and which can be aided through testing.One of these consists of organization and re-organization. It seems true, that not only should things be learned in one way, but they frequently should be learned in others as well, if they are to be really useful. Learning things in other ways means merely making different application of them or of finding new rela-

tionships to them. This is re-organization. Testing can be used to great advantage for achieving this purpose.

7) History tests enchance the intrinsic worth of learning----- Another use of history tests is one which has yet had but little attention, but which is rich in possibilities for both the present and the future. It helps to realize one of the great ideals of education, namely, to enhance for this and for future generations the intrinsic worth of learning.

At a time when the efforts of educators are focused upon an ideal of worthwhile activity on the part of pupils in school, when the curriculum is being scanned to remove traces of artificiality, when courses of study are being pla ned to eliminate subjects or parts of subjects that are included largely because of prejudiced tradition, and at a time when the work of the pupil in school is directed toward making his life there more rich and more like life outside of school, any plan which will enhance the intrinsic worth of learning to the pupil is a step toward the realization of those ideals.

Some learning has passed from the stage of coercion through the stage of reward, past the stage of rivalry, and is now founded on the worth of learning

for its own sake. Tests rightly used, emphasizing the worth of learning, the desirableness of knowledge, especially in terms of its usability, are means to bring more of the materials of education to this level.

We can see the reasons for the ineffectiveness of rewards and sugar-coating as a basis for right learning. We can see the unwholesome ideals connected with emphasizing the rivalry of one's fellows. It should be clear that it is in rivalry with one's best previous efforts, of oneself, that real education results.

In daily living, rivalry with one's fellows seems, from a superficial point of view, to be the paramount motive for success. But one needs only to examine the conspicuous cases of success in his own environment to discover that these are not merely the result of a selfish rivalry with the success of others. The conspicuously successful physician, lawyer, pastor, or teacher has no rival save himself. If this is the criterion for success in life, it should be the criterion for success in school.

Tests rightly used and rightly interpreted furnish a means by which pupils can rival their own best previous efforts. It is a means of promoting one of the

highest types of social education.

8) History tests may measure memory------

History tests may be used to measure the memory of pupils. There may be several types of memory tested. The type of test most frequently used in school is that which determines the amount of material an individual has retained. John has been studying about Africa. The test is one so designed as to determine how much John still remembers of what he was supposed to have learned about Africa. This is a good test of memory of pure knowledge; but, with the emphasis which is now being placed upon other types of memory and upon other attributes of knowledge, the test is, in this respect, somewhat too narrow.

It is frequently good to rememeber things not as separate facts but as related facts. In this case some other type of test is needed: one which not only measures a knowledge of the facts themselves but also an additional knowledge of their relation to one another.

It may also be desirable on occasion merely to have a memory which recognizes a fact, without necessarily having the ability to recall the fact, unassisted. To recognize the truth of a fact, with the fact out of the context in which it may be usually found, constitutes a universal need of daily life and may on occasion

prove as valuable as any other form of memory.

The same may be said of the order of a series of facts. It is often as necessary for an individual to know the order of a series of facts, as to have the memory for the facts themselves.

All these types of memory should be used in school, and we should encourage them, in addition merely to stimulating the remembrance of facts supposedly "memorized" or "learned".

9) History tests may be used for comparison------

One of the results of the testing which is now being done in the schools is that a teacher may compare his class and the work of its individual members with like pupils in like classes in other places. In a nation as large as this - a nation where the ideals of education are (in their more fundamental aspects at least) so universal in all sections, and where the character of the sections and the character of the people differ to such an extent - it is a great value that a teacher should know what may be the standards in other parts of the country, and to be able to compare his standards and his pupils with these. Tests may furnish the only reliable means by which this can be done.

10) Use of history tests to improve teaching-----

History tests may be used in several ways to improve

teaching. One way may be considered in its relation to the pupil. Tests furnish for the teacher a more complete knowledge of his pupils. Through testing he can discover the individual status of the pupils, together with their individual difficulties and misunderstandings. He can also discover the difficulties and misunderstandings that are characteristic of like pupils. When the teacher anticipates these difficulties in his teaching, he is improving it.

A second way may be in the subject matter itself. Tests may be used to increase the value of materials already in use by giving to them wider application, or greater implication, and they may also be used as guides to needed or desirable extensions of these materials. Either of these uses of tests should result in improved teaching.

A third way of improving teaching may be in method. History tests furnish objective results of the methods by which teaching is accomplished. These objective results are not alone a measure of the pupil and his accomplishment; they are also a measure of the effectiveness of the emthod which the teacher has used. When varying methods are contrasted in terms of the results secured, and the better methods are chosen for future use, better teaching is one result.

A fourth way in which teaching may be improved is through the teacher himself. Here the improvement may result from the wider knowledge which the testing makes necessary, or from the greater skill and confidence in the teaching which the tests make possible, or, most important of all, from the added stimulation to constructive thinking which accompanies the testing.

11) Summary-------

History tests are being used in many ways besides merely for the purpose of examination and promotion. They may be used to test the efficiency of a teacher, or to examine pupils for purposes of locating beginning points in teaching, or for determining school status. They may promote review and recall either by increasing the retentiveness of pupils, or by organizing and reorganizing the work that has been covered. They may be used for placement of pupils, or for classification. They may be given in order to diagnose difficulties of pupils, and they may in that way provide a basis for remedial measures. They may be used so as to compare a class or a pupil in one part of the nation with another in a different part, or with the composite pupil of all parts. They may be used for the motivation of school work, for the promotion of real and not artificial interest, bringing with that the hand-

maiden of interest - active attention. They may provide pupils with objective goals for school study. Most important of all, to the extent to which their influence may be felt, history tests may improve teaching.

Chapter 3

Criteria for the Selection of History Tests in the High School.

1) Criteria to be considered.
2) Validity.
3) Reliability.
4) Other things being equal, choose a test that has norms.
5) Other things being equal, choose a test having alternative forms.
6) Choose a test because it is sealed.
7) Ease of administration.
8) Other things being equal, choose a test that is inexpensive.
9) Summary.

Chapter 3

Criteria for the Selection of History Tests in the High School [1,2.]

1) Criteria to be considered-----

It is the purpose of this chapter to present a list of criteria which should be employed in selecting standardized history tests, and to discuss and explain the meaning and application of each. Several of the same criteria are also suitable for use in connection with teacher-made, or unstandardized tests, but others of them have no application in connection with such tests. The reader will have no difficulty in recognizing which are applicable to other tests and which are not, since the discussion, without saying so, explicitly will make clear that certain of them cannot be expected to apply to ordinary classroom tests.

For many workers in educational measurements it is more worthwhile to be able to recognize a good test by its earmarks than to know intimately the many tests - good, mediocre, and poor - that exist today.

The chief criteria or bases of selection, some of which may be subdivided, are as follows:

---

1. From Ruch and Stoddard, Tests and Measurements in High School Instruction pp.45-68
2. From Symonds, Pervical A.A. Measurement in Secondary Education pp.278-307

I Validity
II Reliability
(a) Objectivity
III Norms
IV Duplicate and equivalent forms
V Scaling
VI Ease of Administration
VII Cost

These are arranged in approximate order of importance with regard to the real merit of tests. From the more practical standpoint of selecting history tests which it is possible or practicable to employ, the last two items are often of greater importance than their position indicates.

2) Validity-----

The first concern of one who wishes to choose a standardized history test is to make sure that the test really mesaures what it purports to measure. (Wood[1] emphasizes validity as the fundamental desideratum in test construction). The adequacy and detail with which a test is a measure of a trait, function, or school subject, is called its validity. Validity is measured or determined by the correlation of scores on the test with some independent criterion of the school subject in question. The criterion in the case of history is some other measure of the subject which has no reference to the test whose validity is being considered.

In the case of achievement tests the independent

---

1 Wood, Ben.D. "Studies of Achievement Tests," Journal of Educational Psychology, 17: 125-139. Feb.1926

criteria to be used for validation are few in number, One must usually fall back upon school marks or teachers' estimates of achievement as a criterion. Often where several tests have been constructed in history, it is possible to determine the relative validity of the separate tests by constructing the composite of all the tests, and determining the correlation of each test with this composite.

In judging the validity of a test one should investigate how the material of the test was assembled. Good standardized tests usually contain statements, either in the manual of directions or in articles in technical periodicals, describing how the material for the test was hathered. One should search for such a statement, and, if none is to be found, he may immediately suspect the care used in constructing the test. If there is no description as to how the material was assembled, one may suppose that it was hastily collected much as a teacher collects material for an informal test. With material gathered in this way, a test is almost sure to be short-lived and doomed to disappear because of its inadequate sampling, or because of its irregular emphasis of material.

In the case of achievement tests most of the

validation of the test must be accomplished in the original choice of material of the test. Concerning the principles of the choice of material there is much confusion, particularly in subjects which are now in a period of development or transition. Should the test attempt to measure some typical course as represented by average textbooks or examinations used in the country, or should material representing progressive experiemntal courses be selected? If standardized tests were used solely as measureing instruments, probably the former type would be more desirable, but the truth is that standardized tests also set up in the minds of teachers standards for the selection of material and instruction. For this reason, they ought to represent, in some degree, forward-looking tendencies in curriculum development. Certain principles concerning the choice of materials may be safely given. Five distinct methods or criteria have been employed in the choice of material for standardized tests in hsitory.

1. Criterion of Extrinsic Use

This means that the choice of material is based upon its usefulness. This criterion, however, is difficult to employ unless the studies of frequency

of use have already been made, for it is usually more difficult and tedious to make the studies than to construct a standardized test

2. Criterion of Composite of Textbooks.

One of the most convenient and practical of the methods for selecting the material for a history test is to use a composite of what is given in several textbooks in the subject. This means that the selection of the items in a history test has been based, somewhat, upon an analysis of the contents of several textbooks in the subject, and the construction of the test embodies the best of the materials actually occurring in the books which must be used in history classes.

3. Composite of Requirements in course of Study

Very similar to the use of textbooks is the use of requirements in courses of history. This means that the selection of the items in a history test has been based, somewhat, upon an analysis of the requirements of several courses in the subject, and the construction of the test embodies the most generally required materials.in the several courses.

4. Composite of Teachers' Examinations.

This criterion has been used effectively in some history tests. This means that the selection and validation of the items in a history test have been based,

somewhat, on the analysis and condensation of many final examinations used in all parts of the country over a period of several years.

5. Judgment of Experts.

In all cases, as a last resort, we go back to the judgment of experts. Perhaps we should not say a last resort, because the judgment of experts has been the decisive factor in determining the content of text-books and courses of study.

Selection of Test Items.

Makers of our best standardized history tests have usually assembled at the start several times as many items or elements as it was their intention to keep permanently. It is generally recognized today that a standardized history test should have at least two, preferably more, duplicate forms, so that the test may have its greatest usefulness in actual classroom practice. Often a teacher wishes to re-test her class soon after a first test is given. It is not desirable to repeat the identical test, because certain of the items might be remembered, and, perhaps, especially looked up in the interval. It may be affirmed here that there is less danger of this than is commonly supposed. However, it is usually possible, and hence desirable, to prepare and use tests identical in form

but differing in content from the original test.

The other reason for preparing many more test items at the start than it is intended to ultimately use in the test is in order to discard items which seem to have little validity for the purposes of the test.

Textbooks, courses of study, and teachers' examinations have been used to determine the most valid elements.

3) Reliability-----

The second most important criterion of a history test is its reliability. Indeed it is perhaps as hard to construct tests with the desired reliability as with high validity. If the testing time is of necessity brief, give prime consideration to reliability of the test, and, if the testing time is long, give prime consideration to validity, and but secondary attention to the reliability of the test. To-day the maker of a good test reports the reliability of his instrument in the manual of directions accompanying the test, or in some article of a technical journal which describes the test. One should suspect the reliability of a test whose author has not taken the trouble to determine its reliability.

Reliability refers to the correlation between the results of two forms of the same test. It is the accuracy of a test, - the accuracy with which the test measures whatever it does measure. This is not necessarily what the test claims to measure. Unre-

liability is best exhibited by giving an alternate form of a test and noting the change in score. A perfectly reliable test is one for which the score on a second trial remains exactly the same as on the first trial (barring practice effect). A perfectly unreliable test is one where the scores on two successive repetitions of a test are such as might occur by pure chance - such as one might draw out of a hat containing a number of possible scores. In estimating reliability we are not concerned with repeating the identical test, but in giving a test having the same construction, form, and subject matter as the original test but having different items. In using tests similar in nature but differing in content, one is really comparing the two halves of a test double the size of either form.

Factors Which Influence Reliability.

1. Objectivity.- When the judgment of the scorer enters into the determination of the score, the test is called subjective. A subjective test must be unreliable, for it is obvious that the very same test paper may be scored differently by two different scorers, or by the same scorer at two different times. Objectivity may be determined by inspecting the form of the test and noting whether judgment is necessary in the scoring. If one has to use considerable judgment in determining

the correctness of an item, then the test is subjective. On the other hand, if one correct answer only is admissible for each item, and that one is to be clearly indicated in some way, then the test is highly objective. Objective tests are usually recognition tests, as opposed to recall tests.

The matter of objectivity is one factor of unreliability which can most easily be remedied. Of all the factors entering into unreliability, lack of objectivity is perhaps the most inexcusable, for it is usually possible, by exercising sufficient engenuity, to turn a subjective test into an objective test, with little or no loss in validity to the test.

2. Length of the test.- One of the most potent factors in determining reliability is the length of the test. Wood[1] says that there is a "crucial need for increasing the validity and reliability of our educational measurements. Even our best tests afford only approximately accurate placement of individual students. The needed improvement will be hastened by more care in the construction of individual questions, by drastically lengthening our examinations, and by using a greater variety of appropriate quesion-forms in them."

Two questions give more reliable information of a

---

1. Wood, Ben D.- "Studies of Achievement Tests" Journal of Educational Psychology. 17: 125- 139. Feb.1926.

person's knowledge or ability than one. A test with fifty items gives a more reliable estimate of a person's ability than a test of twenty-five items. A limited sample of anything gives an untrustworthy, unreliable picture of the whole.

3. Evenness of scaling. - Not only should the items of a test be selected because of the fairness and adequacy of the sampling of the topic to be tested, but also because of their range of difficulty. If the items of a test are equally spaced in difficulty, there is no loss in reliability due to coarseness of the measuring scale. Unless one know definitely the difficulty of the items of a test, there is a chance that the items will be bunched in difficulty.

4. Condition of the pupils taking the test. - Other factors of test unreliability have to do with the conditions under which the test is taken, and the variations in the pupils taking the test. With regard to the latter, there is much apprehension. The instability of the individual taking the test is commonly regarded as the outstanding factor in causing test unreliability. Most people feel that strain or nervousness, or a physical condition not quite the best, works havoc with their performance on an examination. Not only does popular impression attribute much of the cause of unreliability to this nervousness,

etc., but even experts in the field of testing make similar statements. But variety of opinion exists here.[1] Kelley believes such phenomena have a certain amount of influence in test reliability.

McCall[2] says, "An automobile horn, the lonesome howl of Jack's dog, the bleating of Mary's lamb, a sudden thought of the swimming hole, growing discomfort of a strained posture, these and a thousand other large and small internal and external influences register themselves in the pupil's scores."

Symonds[3] says,"In a study of reliability, I give evidence to show that the general conditions of an individual exert practically no influence, under ordinary school conditions, on the results of testing. On the contrary, not one of 232 school children showed any appreciable unreliability in tests due to this personal factor."

Symonds[4] says, further, "It is remarkable how, under the stress of taking a test, an individual can pull himself together to do his best. This seems to be the general rule rather than the opposite."

5. Familiarity of pupils with the technique of taking tests. - Much of test unreliability is due to lack of training in the technique of taking a test. For instance, there may be a great variation in the speed

1.Kelley.T.L.-"Note on the Reliability of a Test:A Reply to Dr.Criticism".Journal of Educational Psychology, 15. 193-204, April, 1924
2.McCall,W.A. How to measure in Education, pp.308
3.Symonds,P.M."A Study of Extreme Cases of Unreliability" Journal of Educational Psychology, 15, 99-106, Feb.1924
4.Symonds,P.M.Measurement in Scondary Education,p.295

with which the tests are taken. In tasks which are unfamiliar the pupil would often seem to be struggling with what to do the first time, but, once he knew what was expected he would go much faster the second time, with a clear understanding of the task. Sometimes just the opposite would take place. On the first trial the pupil would commence underlining rapidly, with no clear conception of the problem. Later he would catch on to what was expected, and the new job of actually doing the problems would slow him down.

What Should One Expect In The Way Of Reliability?

One answer is: Expect the highest reliability you can get in the time allowed, - that is, set the time you can afford for testing, and then use that type of test which will give you the highest reliability. Much depends on the number of items that can be asked in the time allowed, and also the validity of these items.

The answer to the question: What reliability should one expect? must remain a qualified one. The factors which enter into unreliability are so many and obscure that tests may fail to show high reliability for unknown reasons.

With thirty minutes of testing, where fifty or more objective questions are asked, we should expect to get

over .80 for a reliability coefficient. With increased testing time, involving a na-ger number of questions, the reliability rises to any reliability desired. One gets in test reliability what one is willing to spend in testing time.

Teachers should come to realize that, in order to make testing worthwhile, the test results should be reliable. This implies continuing the test over two or more school periods. One should aim to get a reliability of .90 in all cases where the results of the testing are to be used seriously in the administrative or guidance functions of the school.

The paucity of determination of reliability is pathetic. The tests for which reliability coefficients or probable errors of a score can be found are a mere handful. No be ter evidence of the youth of the testing movement in the high schools exists than this fact that workers can assemble a test and publish it without demonstration that the test is better than someone's else. The time will surely come when teachers will be trained to demand tests of a specified reliability.

4). Other things being equal, choose a test that has norms.

A standardized history test without norms is practically useless. A norm is usually an average, a medium, some percentile, a measure of variability, or some other

measure derived from one of these. It is most often a mean, or a medium. The term "standard" means a "goal". Norms, or standards, are essential for interpreting the results of a test. The three most common types of norms supplied with high school tests are:

(a) Percentiles (either grade norms or age norms)

(b) Subject norms

(c) T-scores (or other measures based upon the standard deviation or other measures of variability).

A class average should be compared with the norm or average performance of other groups. A pupil's score can only be satisfactorily interpreted in the light of percentile norms, or some similar set of indices, which will permit comparisons of a pupil's performances with a large number of pupils.

Norms should accompany a standardized history test. Most publishers include the norms in the manual of directions accompanying the test, or on the tabulation entry sheet which they sometimes provide.

Select tests for which ther are satisfactory norms derived from a large number of cases. This is evidence that the test has been widely used.

5. Other things being equal, choose a test having alternative forms----- The best standardized history

tests are made with two or more duplicate forms, and test users should insist on using tests that have alternative forms for several reasons.

If it is desirable to repeat a test during the course of the year's work, there will be no danger, if previous test items have been looked up in the meantime and remembered.

Another reason is that oftentimes a teacher wishes to measure his class with greater reliability than a single test affords, and alternative forms provide an easy way of immediately extending the length of a test. The author of a test needs at least two forms to determine the reliability.

When standardized tests were first marketed, authors put the best and most valid material in one test. Now that enough material must be assembled to cover the requirements of two or more comparable tests, the best material is sorted out equally to all the different forms. For this reason, when there are several alternative forms, no one form can contain the most significant, or most valid material. One has to dip down in the reservoir for less important material.

There must always be a mean between the desire to obtain a valid test and the desire to provide for increased reliability by supplying several alternative forms.

6. Choose a Test because it is scaled-----A scaled test is one in which the items of the test progress according to some quality such as difficulty, aesthetic quality, legibility, etc.

A test scaled on the basis of difficulty is one in which each item is more difficult than the preceding one. Most standardized educational tests are scaled on a basis of difficulty.

In order to make a test a true scale in the same sense that scale is used in other branches of science, two other features are essential. One, the items of the test should proceed with equal increments of difficulty. That is, item 3 of the test should be as much more difficult than item 2, as item 2 was more difficult than item 1, and so on. The other qualification is that zero on the scale should represent "just not any" of the thing which is being measured by the test. Difficulty, in the sense in which it is used above, may be determined in two ways: (1) by the combined judgment of a number of competent people, (2) by the decreasing percentage of a group taking the test who can answer successive items. Thorndike[1] has shown that there is a high correlation between these two procedures in determining difficulty. The second method is to be preferred when it is possible.

---

1. Thorndike, E.L., Bregman, E.A., and Cable, M.V.- "The Selections of Tasks of Equal Difficulty by a Consensus of Opinion". Journal of Educational Research, 9:133-139 Feb.1924.

There has been much apprehension in the scaling of tests. In general, the scaling of a test has been much overemphasized in past discussion. To-day,"number of correct items" is nearly universal as a scoring formula. In fact, providing certain general cautions are observed, it is permissible to merely arrange the items of which your test is to be composed, in order of difficulty, without paying much further attention to scaling.

One caution to be observed is that there should be some items on the test so easy that no pupil will get a zero mark. A zero score is called an undistributed score. By this is meant that the scores fail to show individual differences. Two individuals with zero scores may really differ in the ability in question, yet this difference will not appear in the score. A test with zero scores is too hard.

Likewise, there should be some elements in the test of such a difficulty that no pupil in the class will get a perfect score. Perfect scores on a test are also undistributed scores and equally as pernicious as zero scores. Perfect scores fail to distinguish between individuals of high ability. The introduction of more hard elements into a test which yields perfect scores would undoubtedly produce a spreading out of the

perfect scores.

There should be no undistributed scores at any point of the scale. Such undistributed scores sometimes occur in the middle of a test when items are bunched in difficulty.

A final qualification is that the test should be of a general difficulty such that the mean score of a class is equivalent to about fifty per cent of the total possible score on the test.

With these qualifications it makes little difference whether or not the items are accurately scaled for general school usage.

Most standardized tests for use in high school have such a generous time limit that speed is not a factor in the score.

Of course it should be understood that a test which is not thoroughly scaled is useless for experimental purposes (unless highly refined statistical procedures are used).

7). Ease of Administration------Ease of administration should be considered in the choosing of standardized history tests, especially from the more practical standpoint. This criteria is obviously of great importance. It should be judged from two points of view: first, and most important, the clarity of instruction to the pupil, and second, the clarity of the instructions

to the examiner.

Directions to the pupils are properly printed in the test booklets. If the test is broken into parts or sections, each section should be preceded by the directions for that unit, together with samples showing the pupil how he is to indicate his answers. The instructions to pupils should be full, and very simple in phraseology. It must not be assumed that all pupils will holdin mind long and complicated directions.

There is a very simple rule which is observed in good textbook construction, and other instructional materials, viz: Place on the materials intended for the pupil only those printed directions which apply to him, reserving all instructions to the teacher or examiner to a separate place, preferably the Manual of Directions or Examiner's Guide.

A test whose administration cannot be learned by the average teacher in an hour or two is not likely to succeed. At the same time, the beginning examiner must be warned about a multiplicity of small but important details, such as instructions covering the distribution of blanks, the filling in of pupil information data, the observance of time limits, the breaking of pencils, the prevention of disturbing factors, and other requirements of good test conditions.

The directions for giving a standardized history

test in high school should cover the following points:

1. Statement as to what it is a test of.

2. Statement as to how many parts the test has, if the class is allowed to go straight through.

3. Statement as to how many questions there are in a given part of the test, if the class is stopped at the end of each part.

4. Warning to turn the page, if the test continues on the next page.

5. Warning not to dwell too long on any one item or question.

6. A sample exercise which the examiner usually reads aloud.

7. Statement as to the mechanical form of answering - whether underlining, crossing out, writing a number in a parenthesis, etc.

8. Directions as to what to do, in case a pupil finished the test before time is called.

Scoring.

A test should not be primarily selected because it is easy to score, but, at the same time, a good test ought to be as easy to score as possible.

Publishers usually distribute scoring keys with standardized tests. The key should be printed as near to the margin of the paper as possible so that it can be brought close to the pupil's answers. It should be spaced exactly as the material is spaced.

8) Other things being equal, choose a test that is inexpensive----- This criterion does not represent actual test superiority, but is merely a criterion that

answers administrative exigency. Probably no teacher or school administrator needs to be told to select an inexpensive test. There is little or no correlation between the price of a test and its value as a test.

To-day, publishing houses issue series of standardized tests, selected and edited by expert psychologists, print them in an attractive way on good quality paper, and charge prices commensurate with their quality.

Previous criteria are more important than this one of cost, and one should always choose the best test from the point of view of validity and reliability, even if high priced.

What we need in the field of testing is lists of valid elements if known knowledge in each subject, from which a teacher could readily construct her own test, having the optimum validity for her class, and the desired reliability. With such lists available, a city system could construct and print its own tests, thereby freeing itself from dependence on publishing houses, and reducing costs to a minimum.

Test making is still in its infancy, and when workers are through scratching the surface and producing merely usable standardized tests, they will proceed to dig the ground deeper and produce exhaustive lists of elements, together with their difficulty and validity at various stages of learning.

9) Summary------The chief criteria which should be kept in mind in selecting history tests are validity, reliabilty, objectivity, norms, duplicate forms, scaling, ease of administration, and cost.

The two fundamental criteria for selecting history tests are validity and reliability, - all others are subordinate. Objectivity in scoring is one important factor in reliability. Validity refers to whether or not a test accomplishes its purpose, and may be subdivided into curricular, and statistical validity. The former refers to agreement with the content of a desirable curriculum, and the latter to the statistical testing of scores to determine their validity. Reliability is synonymous with accuracy and is determined by giving a test twice to the same pupils. Among the chief factors that affect reliability are objectivity, the length of a test, evenness of scaling of the test elements, and the directions for giving and scoring. Objectivity refers to the quality of a test that there is no doubt as to what the correct answers are.

There is a crucial need for increasing the validity and the reliability of our educational measurements.

A scaled test is to be preferred, although for actual school practice merely placing the items in

an order of difficulty suffices in a comprehensive objective test. For experimental work, however, a test should be accurately scaled. In selecting a test one must be careful to see that it goes to low and to high enough limits.

To-day, the best standardized tests have accompanying them standardized directions and norms, and are provided in two or more alternative forms.

Though not of great theoretical importance, the case of administering a test and its cost are of practical importance.

Chapter 4.

Teaching History through the Use of Oral Questioning.

1) The uses of oral questioning.

2) Oral versus written tests.

3) Conclusion.

Chapter 4.

1) The Uses of oral questioning.

Strictly speaking, oral questioning does not usually constitute an examination. Oral examinations are sometimes employed, but their value for the more serious and final determinationof achievement is doubted. This is no argument against oral questioning. In many ways, the teacher's daily questioning of her pupils is of far more fundamental importance than her final written examination. The point is that oral questioning is more logically a part of initial instruction than of final measurement, assuming that there are at least five roughly distinguishable phases to the complete act of instruction as follows:

1. Initial presentation of materials to be mastered. This phase consists of setting problems to be solved, textbook readings and discussions, teacher's comments on persistent difficulties in learning, etc.
2. Drill to support and fix the temporary mastery gained under the first phase of instruction. This may be drill proper, or it may mean applications and reviews.
3. Diagnostic measurement at the period when phases one and two are thought to be complete.
4. Re-teaching or remedial instruction upon any weak-

nesses revealed under the third phase.

5. Final measurement and evaluation of a more general and less detailed character than that of phase three. This constitutes the final survey of achievement, and leads to a judgment as to whether the individual or class is ready to proceed to new work.

It should be noted that certain of these phases are less prominent than others at times, the relative emphasis varying with the character of pupils, teacher, textbook, subject, motivation, etc.

Oral questioning plays its greatest role in the first, second, and fourth phases of instruction as presented above. It is primarily instructional; its value for measurement is more subordinate. Oral questioning as an art has a long history and a considerable literature. It is worthy of more experiemntal study than it has received to date.

2) Oral versus written tests.

Some reasons why the oral test is used more frequently than the written test in class work are as follows:

1. In the oral test the teacher can check up the errors of individuals on the spot, while in the written test correcting of papers outside of the class period is involved.

2. In the oral test the whole assignment can be covered, while in the written test this is difficult, and teachers feel that they must go over the entire lesson in the class period.

3. Oral questioning is also important in securing the proper sort of attention. Every member of the class should be accustomed to pay such attention at all points in the recitation, that he can take up the work where the one who is reciting has left off. Class control is thereby enhanced.

4. Oral expression is vastly more important than written expression. Training in speaking is invaluable.

5. Requires quick thought.

6. Makes the student more logical in speaking.

7. Prevents misunderstanding of question by student, or of answers by teachers.

Some advantages of the written test in class work are as follows:

1. The written test is generally moreeconomical than the oral test. The reason for this is that, in the written test, all the pupils are mentally active during the entire period of the test, while in the oral test only one pupil is necessarily engaged at a given time. Further than this, by use of the written test, the teacher can test the knowledge and the skill of all

of the members of his class much more extensively than he can by the oral test.

2. The written test gives time to think and organize and present more logically.

3. The student is more at ease.

4. There is less chance for bluffing.

5. It gives all an equal chance on the same question.

6. It gives a chance at more than one question.

7. It is fairer.

8. It is more thorough.

9. It calls for more preparation.

3) Conclusion-----

Strictly speaking, oral questioning does not usually constitute an examination. Oral questioning is more logically a part of initial instruction than of final measurement.

As far as practical, tests for knowledge should be written rather than oral. The written test is generally more economical than the oral test. Student and teacher favor short written tests instead of oral tests.

Probably, frequent short written tests, supplemented by occasional long written examinations, constitute the best form of routine testing.

Chapter 5

Teaching History through the Use of the Traditional Examination.

1) Nature of the traditional examination.
2) Advantages of the traditional examination.
3) Dis-advantages of the traditional examination.
4) Conclusion.

Chapter 5

1) Nature of the traditional examination-----

The rise of objective or new-type examination makes necessary a distinction between the long-established form of test and the more recent and more objective type of examination. The former has come to be known as the "traditional", "essay", or "discussion" examination. This sort of examination has been almost exclusively employed in the high school until the last few years, and still, perhaps, more used than any other.

The traditional examination needs no definition. It is the examination which we all recognize as consisting of five, ten, or more questions, beginning most often with "state in full", "dexcribe", "tell what you know", "analyze", "discuss", "outline", "explain", "summarize", etc. The pupil is free to write what he chooses as a response to the stimulus question. It is to be contrasted in its mechanics with the newer objective examination in that the latter calls for underlining, crossing out, checking, etc., instead of discussion.

The traditional examination cannot be scored mechanically by keys or stencils, but must be evaluated or scored subjectively.

Within the last few years a considerable amount of

space and time has been devoted to condemning traditional examinations, and to showing or attempting to show that those of the new-type are superior. Many valid points have been made, but frequently, perhaps usually, the merits and advantages of traditional examinations have been overlooked.

All too often it has either been explicitly stated, or else implied by the trend of the discussion, that the abolition of traditional examinations is desirable. It is very unfortunate that such an attitude should have been taken and expressed. It was only natural that the protagonists of new-type tests should, in their enthusiasm, over-estimate and over-state their value, but in many instances this has been done in such extreme fashion that little excuse therefor is apparent.

A thoughtful consideration of the question will, undoubtedly, lead to the conclusion that each of the two types above has its peculiar points of strength and its distinct advantages in actual use. Therefore, it is believed most emphatically that a complete testing program of any teacher should include some use of both kinds. In other words, it is not a question of deciding whether the essay, or the new-type examination is the better, and then of making exclusive use thereof, but rather of determining the

occasion and circumstances under which each is most valuable, and then employing each in accordance therewith.

It is the purpose of this chapter to outline clearly, and also as completely as possible, both the advantages and the dis-advantages of the traditional examination, and to arrive at some sort of a conclusion at the end of the chapter.

2) Advantages of the traditional examination-----

The traditional or essay examination is easier to make, especially as regards the amount of time required. The comparitively few questions of this type needed for an examination can be made in much less time than the relatively large number of new-type items. Moreover, the fact that essay examinations can almost always be written on the blackboard, or even given orally, with satisfactory results, makes it but little trouble to employ.

At present, most teachers are more familiar with traditional or essay than new-type tests and, therefore, are better qualified to make and give them. This condition is rapidly being changed, however, and will not long be a valid argument for their use.

It is true that it does not require a great deal of study on the part of a teacher to acquire a fair knowledge and understanding of the new-type examina-

tion, but many teachers have been unwilling to put forth the requisite amount of effort, even when their attention has been directed along this line. Until the many teachers of this sort are better trained and informed about new-type tests, it is probably unwise to attempt to compel, or even induce, them to make a large use of such tests.

It is probably true in practice that essay-type tests encourage less guessing. It often allows a pupil considerable degree of freedom in choosing the form of his answer, and not only allows, but even requires, him to select from a fairly large stock of information the portion which he wishes to use. Moreover, the situation is frequently such that many items which might be selected are neither absolutely right nor wrong. Thus, judgment and other qualities are called into play much more than on most new-type tests.

Thus, essay examinations reveal certain facts concerning individual differences in the quality of mental activity which are not shown by those of the other variety.

In fact, it has been claimed that the one most significant merit of traditional or essay examinations is that they appear to test certain desired outcomes and mental processes, except memory, better than do

the other kind. Although new-type exercises may measure originality, initiative, power to organize, to interpret, to analyze, and synthesize, and various other reasoning processes, to some extent, they appear to do so much less thoroughly and satisfactorily than wll-made discussion questions. In other words, the latter possess higher validity for these purposes than do the former. Monroe[1] say, "traditional examinations call for the functioning of distinct types of mental ability not demanded to so great a degreein any other kind of school work, and should not, especially in the case of final examinations, be abolished, nor replaced entirely by any other form of shhool exercise. Teachers and pupils, more and more, should be impressed with their unique educational value."

They also provide opportunities, such as cannot be afforded by tests to which the answers are single words or phrases, for measuring pupils' power to express their thoughts, to write well, to use correct English, and other related abilities and habits.

The essay examination, if properly administered, not only measures ability to organize and express ideas, but also gives training in such ability. Even though it ia not admitted that it is an important

---

1.Monroe,W.S.- "Written Examinations and Their Improvement". The Historical Outlook, 14: 306-318. Nov.1923

function of an examination to give such training, little if any objection can be raised to the incidental benefits of this sort which it can be made to yield.

It is also claimed that the discussion-type of examination can be more easily and directly adapted to various kinds of subject-matter. This advantage should not be over-emphasized since the large number of varieties of objective tests renders possible their adaptation of many kinds of subject-matter, but still appears, on the whole, to be present.

The claim has been made that the ordinary discussion-type of examination encourages less dishonesty on the part of pupils because it is much harder to cheat upon them. There is some justification for this claim, because the answers in the new-type examination are sho short that it is relatively easy to see the answer of a neighbor by a hasty look. Not infrequently, however, in the case of an essay examination sufficient information can be gained by a mere glance to enable a pupil to profit thereby. Nevertheless, on the whole, it cannot be denied that there is some validity to this argument,

3) Disadvantages of the traditional examination------

The traditional variety of tests are less reliable than those of the new type. A considerable mass

of data, concerning which more will be said a little later, has been accumulated in support of this statement.[1]

There appear to be two chief causes for the difference in reliability. One of these is that the typical essay examination contains comparatively few separate questions on exercises and that hese are too few in number to constitute a satisfactory or reliable sampling of pupils knowledge or achievement. Too few topics are covered, and there is too much chance that these few will be among those which some pupils just happen to know, and others not to know. This defect could be remedied by including a much larger amount of questions, but so doing renders a traditional examination of too great length.

The second of the two chief causes for the lower reliability of traditional examinations is that they are subjective in their scoring. The answers to most questions of the traditional type cannot be scored as definitely right or wrong, but may be partially right to almost any degree. As a result, great differences of opinion exist among supposedly competent teachers as to how much credit shall be allowed for the same answers. Odell[2] cites evidence showing the unreliability of teachers' marks has consisted of, or been based upon, the marks given traditional examina-

1. Odell, C.W., Traditional Examinations and New-Type Tests. p. 183
2. Odell, C.W., Traditional Examinations and New-Type Tests. pp. 5-7

tion papers, and therefore tends to prove the point just made. Starch and Elliott[1] show that, among the chief results of a series of investigations relating to the reliability of grading work in an essay-type history test that "the marks assigned to the same paper by different teachers vary enormously, in fact much more widely than the average teacher would anticipate."[2] Douglass tells of an experiment in which twenty-eight American History examination papers were marked by four high school teachers of history, and then re-marked by the same teachers several months later. The average differences in marks given the same paper by the same teacher were greater than five per cent in the case of all four of those doing the marking. In one instance there was a difference of twenty-five per cent between the two marks given the same paper by the same teacher. The examination used was, of course, of the traditional type.

The unreliability in scoring responses to essay examinations, due to teachers' subjectivity, is only in part caused by disagreements among yeachers as to just what the correct answers are. Much of it also results from the fact that teachers do not agree as to the relative importance and therefore the weighting

---

1. Starch, D., Elliott, E.C., - "Reliability of Grading Work in History". School Review 21: 676-681. Dec. 1913
2. Douglass, H.R., Modern Methods in High School Teaching. p.368

of the different parts of the examination. Some teachers attempt to weight according to supposed difficulty, others according to the importance of the facts, or of the mental activities called for, and others on still different bases. Not only do their judgments differ as to how difficult the various exercises or questions are, but also as to how great weight should be assigned particular questions, upon the relative difficulty of which they are agreed. Most teachers who determine weights on this basis count more upon the more difficult questions, on the ground that hreater ability is required to answer them correctly. Some, on the other hand, count more heavily on the easier questions, because they believe that it is a greater dis-credit to pupils to be unable to answer these, and that, therefore, they should be penalized more heavily, if they fail to do so.

Still further, teachers are influenced in the marks which they give pupils written work by the pupils' past records, by the general opinion which they have of the quality of their work, by handwriting, neatness, language, usage, style, and so forth. In many cases, teachers are unconscious that they are so influenced, but, nevertheless, the condition is very real, and almost impossible to avoid. Moreover, the

merit of a paper as a whole is likely to influence the marks given separate questions. If the answers to the first few questions have been very good, the marker is liable to rate any later poor answers too high. Similarly, if the first few answers have possessed little merit, later good answers are liable to be discounted. It has been related that two students in an English training college, named respectively Smith and Jones, were close friends. They were both members of the same English class, and as such had to hand in essays every fortnight. They consulted one another in their work, exchanged ideas, and produced essays apparantly very similar. The first essays were rated with a mark of "Very Good" upon Smith's, and of "Very Fair" upon Jones', and second, third and succeeding essays received the same marks, except that Smith occasionally had his raised to "Very Good Indeed", and Jones had his reduced to only "Fair". On one occasion they exchanged and copied each other's essays, that is, Jones handed in the essay really written by Smith, and Smith that written by Jones. Nevertheless, each on the essay not his own received the same mark he had been receiving on his own work. Evidently, the instructor had the firmly established idea that Smith could write essays of considerable merit whereas

Jones could not.

A most striking illustration[1] to show the highdegree of unreliability of marks given traditional examination papers is as follows: About half a dozen expert readers were marking a set of history papers. One of these readers for his own convenience prepared what he considered a model paper, that is, a paper containing supposedly correct answers to all the questions. By accident this model paper was included with the students' papers and passed on to several other readers. They rated it on the supposition that it was a stuednt's paper, and assigned marks to it varying from forty to ninety per cent.

Even when teachers are aware of the large element of variability commonly present in their marks and endeavor to reduce it by careful marking, they are unable to do so to a satisfactory extent. Some reductions[2] undoubtedly can be made, but Ruch, as well as others, offers experiemntal evidence which shows that even teachers who are aware of the larger scources of unreliability and endeavor to avoid them, still disagree markedly, in a few cases even over fifty per cent, as to the marks to be assigned papers.

A disadvantage of the traditional examination connected with unreliability, but yet worth mentioning

---

1. Wood, Ben D., "The Measurement of College Work" Educational Administration and Supervision, 7: 301-334 Sept.1921
2. Ruch, G.M., The Improvement of the Written Examination. pp. 55-62.

separately, is that pupils often realize that the marks they receive are to some extent due to chance, or other causes, which should not be operative. Not infrequently when pupils compare papers after receiving them back marked, they find that answers containing almost exactly the same material have received different numbers of points, or even that an answer containing more of the facts called for than does another has been marked the same, or sometimes lower. Pupils feel, therefore, that many of the ratings they receive are really unreliable and unjust. On the other hand, the comparitively high objectivity of new-type tests produces a distinctly favorable re-action.

An essay examination may reveal a general lack of knowledge on a certain topic, but it rarely points out the exact points which need attention. New-type examinations point out vert definitely the particular things which are not known, and thus pave the way for very definite and purposeful remedial instruction.

To some extent traditional examinations discourage systematic and worthwhile review. Since they generally touch only upon a few topics out of the large number included in a course, pupils are likely to take a chance that the few which such tests cover

will happen to be among those that they think they know fairly well. Sometimes pupils do not do this, but try to guess what topics will be dealt with on the examination , and then review intensively on those, neglecting all others. Others make little or no attempt to study because they think it too much a matter of chance whether not doing so will aid them materially in responding to the examination questions.

Another disadvantage of the traditional examination is that it does not test the achievements of pupils whose powers of expression are poor. That is to say, because of difficulty in organizing their thoughts and expressing them in clear language, pupils may know more about the question or topic to be discussed than they indicate in their responses. Thus their answers depend to some extent upon their ability and knowledge along other lines than the subject being tested. It is certainly desirable to test ability in expression, but it should not be done in such a way that a pupil's marks in history or algebra, for example, is a compound mark expressing a mixture of his achievement therein, and also in language, with the proportions of the two which enter into it, unknown.

More or less similar to the disadvantage just

mentioned is that traditional examinations frequently test speed of writing to an undesirable extent. Some pupils' rates of writing, and freedom from fatigue while writing, may be enough greater than those of other pupils, whose actual ability and achievement in the subject being tested is the same, to make very material differences in the marks which they receive upon their examination papers. This can easily be avoided by allowing sufficient time for all pupils to finish, but, as will be shown later, doing so often leads to certain other undesirable results.

Another limitation likewise closely connected with those just stated is that too much of the time spent in answering essay examination questions is ordinarily devoted to what may be called the mere mechanics of answering. That is to say, the act of writing, and the determination of the form of answers consume much of the pupil's time and attention which should be devoted to real thinking about the questions asked.

Because of the fact that language and handwriting abilities play such a large part in pupils' answers to discussion examinations, and further, because if sufficient time is given to these matters, there is frequently not enough time left to devote the proper amount of attention to the subject-matter itself,

pupils tend to develop hurried and careless habits of expression and writing. They hasten to put down on their papers whatever pertinent facts they know, and pay little attention to the form in which they are expressed. This effect is rendered still worse by the fact that many teachers, especially high school teachers of other subjects than English, pay little attention to such matters as spelling, composition, punctuation, sentence structure, quality of handwriting, and so forth. Even if they do correct mistakes along these lines, the attention of the pupils is not usually called to these corrections in such a way as to make them very effective. Because of these facts it may even be said that many essay examinations give positive training in ungrammatical and un-rhetorical expression, poor handwriting, and other undesirable habits.

It is practically impossible to gibe essay examinations so that they test the rate of a pupil's response or thinking in the subject dealt with. They frequently serve to test rate of writing, but not rate of mental activity. It is not necessary, or even highly desirable, that all examinations should measure rate, but it would be very unfortunate if none did so. In practically every activity outside of school life, the person who can perform a task

as well as another, and do so in less time, is rated as more efficient, and the same should be true in much of the rating of school pupils and their work. In actual practice, however, essay examinations as administered and marked have frequently tended to produce the impression that correct answers are equally valuable whether given in a short or in a long time.

One point in which traditional examinations possess considerable dis-advantage over new-type ones is in the difficulty of scoring. Careful and accurate scoring of the answers to traditional questions is relatively difficult and requires considerable time. In the case of many already over-worked teachers the result is that this added burden, in addition to their other duties, is sufficient to result in a lowering of their general physical and mental vitality, and therefore of their teaching efficiency. Some teachers avoid this result by reducing the number of examinations below a desirable minimum, and others by scoring pupils' responses so hurriedly and carelessly as to lose many of the possible benefits to be derived from giving examinations. It is also frequently possible for pupils to deceive themselves as to the correctness

and quality of their answers in discussion examinations even though the papers have been well criticized and marked by the teacher.

It is usually difficult, if not absolutely impossible, to secure satisfactory norms from essay examinations. The difficulty of doing so places very decided limitations upon the possibilities of comparing achievement in different classes or groups of pupils, and thereby renders it harder for teachers and others to learn whether the achievements of their pupils are equal to what they should be or not.

Because it is comparatively easy to dash off a few discussion questions in almost as short a time as is required to write them, many teachers fall into the habit of doing so, and of giving little or no thought to the selection and formulation of the questions and exercises employed. As a result, important topics and portions of the subject-matter studied are often entirely or almost entirely neglected, whereas others are dealt with much more frequently than there is any need for. Hastily-made questions are liable to be poorly worded, obscure, and indefinite, with the result that teachers in scoring either oenalize pupils who cannot understand the questions, although the fault is thie own, or else give credit for answers which are really not

what was wanted. Furthermore, such careless formulation of questions serves to increase the unreliability of marks because of securing poorer samplings of the subject-matter covered.

It is frequently fairly easy for pupils, especially those of more than average intelligence, to bluff on essay examinations. A pupil may know nothing or practically nothing of what is actually called for by a particular question, but if he has some knowledge of the general topic with which the question is connected, and perhaps also some skill in guessing, he can frequently produce an answer for which he will receive much more credit than he deserves. This is especially true where the teacher is marking the papers hurriedly and carelessly. She is liable to notice that the pupil has written an answer of considerable length, and that it contains a number of words and expressions which have something to do with the topic, and therefore, without careful examination of what is written, give him a fairly good mark upon it.

Some of the claims made for traditional examinations are not fully valid. Frequently the time limits upon such examinations are so short that they really become chiefly tests of memory rather than of reasoning, organization, and other abilities, even though

they might provoke activities of these sorts, if sufficient time was given. Moreover, the actual exercises or questions which they containrrequently deal with actual mat rial to just as great an extent as do those in ew-type tests.

Still further, several varieties of the mew-type tests do stimulate and measure pupils critical ability, their discrimination, judgment, and so forth. Such types as the true-false, which requires them to decide whether statements are true or not, the multiple-answer, in which one, or sometimes more, correct answers must be selected from a group, and others can be made to serve these purposes.

Even though new-type tests deal largely with separate points or facts, the material covered can, if desired, consist of general principles, rules, laws, and so forth, as well as mere bits of fnformation.

Therefore, although it is probable that traditional examinations do test a greater variety of mental processes than do objective tests, it is not inevotable that they do so, and the latter also can be made to measure, at least to some degree, most of these processes.

4) Conclusion---------- Although there has been much recent unfavorable criticism of traditional examinations, they should not be entirely discarded in favor of new-type tests, but should be used on some occasions. Each of the two general types just men-

tioned has its peculiar merits and advantages, and should be employed when it best fulfills the desired end. Traditional examinations are usually much easier to prepare, test a number of mental processes better than does the new-type, do not offer a great opportunity for guessing and perhaps not for cheating, are not as liable to the danger of confusing the pupil, and in several minor ways are to be preferred. On the other hand, traditional examinations are, as is hown by a considerable mass of evidence, less reliable than the new-type tests because they do not secure as good samplings of pupils' ability and knowledge, and because their scoring is relatively subjective and more difficult. Bluffing is also less difficult.

Chapter 6.

Teaching History through the Use of the Standard Test.

1) Growth of the standard test.

2) Nature and purpose of the standard test.

3) Additional advantages of the standard test.

4) Some limitations of the standard test.

5) Example of the standard test.

6) Conclusion.

Chapter 6

Teaching History through the Use of the Standard Test.

1) Growth of the standard test[1] -------- Standardized examinations have just completed the first quarter century of their existence. From a few pioneer attempts Rice, Thorndike, Stone, Courtis, and others in the fields of spelling, arithmetic, and reading, the movement has grown until conservative estimates place the total number of available tests and scales at at least five hundred; thereare probably considerably more. It is impossible to secure even approximate estimates of the numbers of standard tests administered annually. There are several educational tests whose sales have passed the million mark annually. In one or two cases, two million is a more nearly correct figure. The total number of standard tests sold during the past year (1928) is probably at least twenty million, possibly somewhat more.

These figures, estimates as they are, point to the importance of standard tests as measures of the results of teaching. It seems certain that the curve of the use of standard tests is rising more rapidly than is that of the increase in school popluation.

A tendency now is that the use of standardized tests is no longer commonly thought of as an experiment, or as something definitely apart from regular

1. From Ruch, G.M. The Objective or New-Type Examination. pp. 21-22

instructional and other activities. Instead, it has come to be an integral part of the work of teachers in many systems, both large and small. Undoubtedly, the most notable example of a large city system in which this is true is Detroit, where an extensive program was developed, but in hundreds of others it has a prominent place. Some large cities, such as Philadelphia, Detroit, and Denver, construct many of their own standardized tests, instead of purchasing them.

2) Nature and purpose of the Standard test-----

In the first place, so-called objective tests may be divided into two groups, standardized tests and the new-type examination. Originally and in its narrowest sense "standardized" was applied to a tests that had been widely enough given that the results therefrom indicated what might be expected of pupils of a given age, grade, or other homogeneous group.

In general usage, however, the adjective "standardized" or "standard" also implies that the test in question has been carefully constructed according to certain general principles, and embodies exercises of such forms that pupils' responses are relatively, if not absolutely, objective.

Furthermore, practically all tests which merit the

name standardized are commercially available, that is, may be purchased from a publisher by anyone desiring to do so.

The essay examination suffers from one major defect not inherent in the standard test or the newer objective examination, viz: experience and experiment have shown that the results of an essay examination cannot be evaluated fairly by human minds. Its inaccuracies are those of the human mind and the human prejudice. Such examinations seemingly cannot be freed from the personal equation.

To a degree that an examination mark or grade reflects the knowledge, attitudes, and prejudices of the marker of that examination paper,- the examination is not a true measurement, since all are surely agreed that it is the accomplishment of the pupil which is to be measured. If the same pupil's paper is graded all the way from 40 to 90, (as many investigators have found [1]), there is but one conclusion to be drawn, viz: the pupil has not been measured. To be at the same time a "40" pupil (a dunce) and a "90" pupil (a candidate for the class validictorian) is not only unthinkable but palpably untrue! Such a finding raises the suspicion that he is neither, a conclusion that can well be supported on

---

1. Ruch, G.M., The Objective or New-Type Examination. p.20

the ordinary logic underlying our basic theorems of possibility.

To be taken at face value, any examination result must meet many stringent criteria, and one of these is that it is a measure of the pupil - not the teacher, not his class, and not the school system. Yet it must be admitted by any fair minded student of the literature that the traditional examination is prone to tell us as much, or almost as much, about whom the pupil had for a teacher, as it does about the educational equipment of the pupil himself.

The technical term for this weakness in the common essay-type examination is, in our modern educational terminology, subjectivity of marking. It was as a relief from this admitted weakness that the standard test and the objective examination were introduced.

The standard test was introduced to serve several purposes. These serve to orient our thinking about tests and examinations in general. The principal aims of the standard test may be listed as follows:

1. They (as the name implies) represent an attempt to control or standardize the conditions of the examination period, with respect to directions, time allowances, method of responding, etc.
2. They are objective or impartial, i.e., the personal equation of the examiner is minimized or eliminated -

minimized in the administration, and eliminated almost or quite completely in the scoring of the examination.

3. They provide norms or standards (as the name further implies) by which the scores of individual pupils may be evaluated and interpreted in the light of facts. Such facts are the performances of large numbers of supposedly typical pupils on the same tasks.

These aims can all be attained to degrees commensurate with the practical needs of education, the third aim being the most difficult, and, on the whole, decidedly the least important.

One characteristic of standard tests is their comprehensiveness within the field measured. It is constantly being emphasized by the makers of standard tests that justice in comparisons of two or more communities on the basis of these tests depends upon the extent to which they are comprehensive, - the extent to which they apply equally to the needs of the different communities which they are designed to serve. A lack of a given content in a test would re-act against a community which emphasized that group of elements, and might act positively for that group which neglected it.

Standard tests are so constructed, through the

selection of the elements which they contain, that they may be as comprehensive within the field to be measured as it is possible to make them.

Another characteristic of standard tests is the universality of their content. A field of teaching which is localized or provincial in character cannot be placed successfully in a standard test. Theoretically, such a test could be standardized for a given restricted community where the localized subject matter was used, but actually the success of standard tests lies in their almost complete use of universalized subject matter. A test in arithmetic, for example, should utilize the subject matter in arithmetic of admittedly universal teaching. Arithmetic is a unit of subject matter that is widely distributed in much the same way, and is, therefore, peculiarly adapted for use in standard tests. Other subject matter in the schools - geography, and history, for example - have far less universality of content, and are, therefore, to that extent restricted in the values of standard tests which can be utilized for them. Rugg[1] is convinced "that our judgment and grading of pupil re-action will, at least, be refined through the use of standard history tests. They are valuable, first, to check the basic aims and outcomes of history."

1 Rugg, E.M.- "Character and Value of Standardized Tests in History". School Review, 27:757 - 771. Dec.1919

3) Additional advantages of the standard test-----

In general, standardized tests are more carefully and scientifically made than non-standardized ones. Their authors are usually better versed in methods of test construction, and have a wider knowledge of subject matter than do regular classroom teachers. Therefore, standardized tests conform more closely to general or best practice than ordinary examinations, and thus tend to produce uniformity. Since more time and care is devoted to their construction, they are generally more objective, reliable, and valid than non-standardized tests, although the latter can be made equal to them, if sufficient pains are taken to do so.

The distinguishing feature of standardized tests, that norms have been established, (previously referred to), is an advantage in that it renders possible the comparison of pupils with others. Norms are, however, frequently too general to be of high value, and are even liable to misinterpretation and misapplication because of this fact. Also many tests may serve their purposes without such comparisons being needed.

A third advantage of standardized tests is that their use saves time in both preparation and scoring. The objection may be raised, however, that time devoted to thoughtful preparation of test exercises is

profitably spent and should not be lessened. Furthermore, the saving of time in scoring is little, if any, when they are compared with the best new-type tests made by teachers.

A fourth advantage of standardized tests is that most good standard tests possess two, or occasionally more, equivalent forms, thus enabling one to test the same abilities of a group of pupils two or more times with tests of the same difficulty. This makes it possible to measure progress much more accurately than if such instruments were not available. It is difficult for teachers to prepare such duplicate forms in a manner that insures equivalence.

4) Some limitations of the standard test[1]----

1. Standard tests are inflexible and cannot be closely adapted to the idiosyncrasies of local school conditions. They are, of necessity, general enough to meet moderately well a wide variety of curricula.

2. In view of the foregoing, they need constant supplementation in a complete measurement program.

3. Standard tests are somewhat expensive. The range of prices varies from about one cent per pupil to at least ten cents per pupil. This, of course, is a practical limitation, not a theoretical one. It should also be noted that there is considerable correlation between cost and worth. As is the case with

---

[1] Ruch, G... The Objective or New-Type Examination. pp. 22-23

all commercial products, tests are sold in a competitive market, and costs are reckoned upon the basis of the expenses of production.

4. The majority of standard tests are of little value. A large number are nothing more than "examinations with norms", produced by persons without special training or knowledge of test construction. If one hundred of the best were selected and the rest destroyed, the loss would be negligible.

Only the first mentioned of these limitations of the standard test is serious. The others may be overcome by careful selection, by the planning of measurement programs, and by efficient school budgeting.

It would appear to be impossible to adopt the standard test as the sole element in a measurement program. It might well repay the cost, but it is to be doubted whether local needs could ever be met satisfactorily. Both the traditional and the new-type examination are free from this limitation of non-adaptability to local school curricula.

Another critic, Tryon[1] says, as regards the limitation of the standard test, that "history test-markers have had difficulty in devising exercises that test a variety of mental processes, such as reasoning, association, and comparison. Too many of the exercises now available test memory only".

1. Tryon, R.M.- "Standard and New-Type Tests in the Social Studies". 18:172-178 April.1927

5) Example of a standard test[1]-------

Part I

Directions. Read each of the following statements very carefully. If a statement is true, place a plus (+) in the parenthesis following it; if it is false, place a (o) in the parenthesis following it. If you are not sure whether a statement is true or false, leave the parenthesis blank. Do not guess.

Twenty minutes.

Samples.

a. George Washington was the first president of the United States ( )

b. The panic of 1857 had no affect upon the economic life of the South (o)

---

1. After the passage of the Toleration Act of 1649, Maryland enjoyed greater freedom of conscience in religious matters than did Rhode Island ( )

2. Horse-racing, corn-husking bees, and card playing were forbidden by law in Colonial America ( )

3. The Proclamation of 1763 forbade colonists to settle in the territory acquired by Great Britain from France as a result of the French and Indian War ( )

4. The economic development of Pennsylvania was retarded by William Penn's refusal to admit German immigrants ( )

5. The principal motive of the majority of those who left Europe for America during the Colonial period was to better their social and economic conditions. ( )

If you finish before the time is up, go on to Part II.

Number right ..... Number right.....
Number wrong ..... Number wrong.....
Number omitted..... Right minus wrong.....(Score)

---

1. Excerpts from Columbia Research Bureau American History Test. Published in 1926 by the World Book Co., Yonkers-on-Hudson, N.Y.

Part II

Directions. Below are eight groups of items, each of which is divided into two columns. Each item in the left-hand column is numbered. Each item in the right-hand column is followed by parenthesis. Place in the parenthesis the number of that item in the left-hand column that is associated with the item in the right-hand column. Each group is a separate problem; do not match items in different groups. Twenty minutes.

Samples.

a. 1. 1492 Declaration of Independence.....(3)
2. 1620 Discovery of America.....(1)
3. 1776

I. 1. Pennsylvania Platoon system.....( )
Complete religious toleration.....( )

2. Massachusetts. Nathaniel Bacon.....( )
Cod-fishing.....( )

3. New York. Duke of York.....( )
Governor Berkley.....( )

4. Virginia. Largest colonial city.... ( )
James Oglethorpe.....( )

5. Georgia. James Otis.....( )
College of William and Mary( )

II. 1. 1620 British Navigation Act.....( )
2. 1643 New England Confederation formed.....( )
3. 1660 Invention of cotton gin.....( )
4. 1763 Lewis and Clark Expedition...( )
5. 1793 End of French and Indian War.( )
6. 1804
7. 1825

If you finish before the time is up, complete Part I or go on to Part III

Number right.....(Score)

Part III

Directions. Below are several statements and questions, each of which is followed by five

phrases. Mark in the parenthesis the number of that phrase that correctly completes the statement or answers the question. (One, and only one, phrase is correct in each case).

Thirty-five minutes

Sample:
a. One of the principal products of Colonial New York was _____

1.Rice 2. Indigo 3. Flour 4. Gold
5. Aluminum.....(3)

---

1. By the Molasses Act of 1733 England sought _____
   1.to drive the colonial rum manufacturers out of business.

   2. to safeguard the interests of the British West Indian planters.

   3. to ruin Dutch shipping

   4. to bring on a war with Spain

   5. to encourage the slave trade..... ( )

2. The Shenandoah Valley was principally settled by _____

   1. the Dutch 2. Irish Catholics
   3. German and Scotch-Irish
   4. Scandinavians 5. Huguenots.....( )

If you finish before the time is up, complete Parts I and II, or go on to Part IV

Number right..... (Score)

Part IV

Directions. In each of the blanks at the right put the word or shortest phrase that will complete the sentence correctly. Write carefully and clearly.

Fifteen minutes

Sample:
a. The name of the first permanent English settlement in America was..... Jamestown)

1. The name of the author of the pamphlets entitled "Common Sense" as_____ ___(_____)

2. The system by which natives endeavored to sell much and buy little, in order to secure a so-called "favorable balance" of trade was known as _____(_____)

If you finish before the time is up, look over all four parts and correct any mistakes you have made.

Number right.....(Score)

6) Conclusion---------Standardized tests have a legitimate place in high school measurement in history and should not be wholly displaced by other types of measurement. Certain advantages accrue from the standardized tests which the others cannot supply.

They are characterized by the care exercised in their making, by the objectivity of their scoring, by the comprehensiveness within the field measured, by the universal character or their content, and by the fact that they are accompanied by norms, or standards of achievement.

Standardized tests are generally more objective, reliable, and valid than either the traditional-type, or the new-type examination. Also, their use saves time in both preparation and scoring. Equivalent forms provided by most good standardized tests insure greater accuracy in measuring progress.

On the other hand, the standardized test should not

be the sole element in a measurement program. It has its limitations as well as its advantages. Among the former is the serious one that standardized tests are not sufficiently flexible to satisfactorily meet local needs. They are non-adaptable to local school curricula.

History test-makers have had difficulty in devising exercises that test a variety of valuable mental processes. Too many of the exercises now available test memory only.

Chapter 7

Teaching History through the Use of the Objective or New-Type Examination.

1) Nature of the new-type examination.

2) Advantages of the new-type examination.

3) Dis-advantages of the new-type examination.

4) Five most generally useful varieties of the new-type examination forms.

5) Conclusion.

Chapter 7

Teaching History through the Use of the Objective or New-Type Examination.

1) Nature of the new-type examination------ The new examination or, better, the new-type test, is the name commonly given to tests or exercises, generally constructed by a teacher for her own use, that make use of the forms and scoring methods of standardized tests, so as to possess relatively high objectivity, but have not gone through a process of careful trying out of material included, have not been given to large numbers of pupils, and are most generally available for use by others.

The new-type tests include true-false statements and yes-no questions, single-answer questions, multiple-answer exercises, matching exercises, completion statements, and other similar types.

No sharp distinction can be drawn between the standardized test and the new-type test, since there are tests in all stages of development, from an ordinary new-type test constructed by a teacher for use with a single class,up to thoroughly standardized test. Both are thought of as opposed to traditional examinations, the primary difference being that they call for very brief pupil responses, and are objective

or nearly so.

The aim of this chapter is to state as fully as possible both the advantages and the dis-advantages of the new-type examination, to illustrate the varieties of the new-type examination forms which are most valuable and which, perhaps, receive the widest use; and, finally, to form a conclusion of the entire chapter.

2) Advantages of the new-type examination-----

The merit of new-type examinations which is probably most often stated first is that they are more reliable than those of the traditional variety. New-type tests permit pupils to respond to a great many more items or exercises than do traditional examinations consuming the same amount of time. Therefore they yield much better and more comprehensive samplings of pupils' achievement, and so result in more reliable marks. It is very unlikely, when a pupil must respond to a large number of exercises, each of which calls for a response more or less distinct from that of any other, that he will just happen to know, or not to know, a much larger proportion of then than is true for the total amount of subject-matter covered by the test.

The second of the two chief causes for the higher reliability of new-type tests is that they

are objective or nearly so in their scoring. The answers to most questions of the traditional type cannot be scored as definitely right or wrong, but may be partially right to almost any degree. As a result, great differences of opinion exist among supposedly competent teachers as to how much credit should be allowed for the same answers.

Some rather convincing evidence concerning reliability is presented by Ruch[1], who is one of the leading advocates of the new-type. In one place he gives figures showing the reliability coefficient of eight New York Regents' Examinations in their ordinary or subjective form, and likewise of the same examinations when converted into objective form. The average coefficient of reliability in the second case, that is, for the objective form, was .65, whereas the average for the subjective form corresponding to this was only .42. If a correction were applied to balance the fact that pupils spent more time working on the subjective than on the objective form, the figure of 165 should be raised to .69. The reason or justification for making such a correction is as follows: If a test is lengthened by adding more of the same type of exercises as compose the original portion, and other conditions are in no way changed, its reliability is increased. This

1. Ruch, G.M., The Objective or New-Type Examination. pp.23-53

increase is due to the fact that making it longer causes it to yield a more satisfactory sampling of the total field covered. Such an increase in the length of a test results in increasing its reliability by a ratio equal to the square root of the ratio of its length after the additional exercises have been added to what it was in the first place. For example, if enough similar material is added to a test to make it four times as long as it was originally, the reliability of the lengthened test is twice as great as that of the first one, since the square root of four is two.

It would be easily possible to quote dozens, probably even hundreds, of reports of results which agree with those of Ruch, that is, show greater reliability for new-type than for traditional examinations. There are, however, a few cases in which data have been obtained which indicate an opposite conclusion. Thus Crawford and Raynoldo[1] conclude from their experiments that fifteen out of twenty comparisons indicate that traditional examinations possess greater reliability than true-false tests. They state, however, that the true-false tests used were made by persons comparatively unskilled in so doing, also that the students upon whom the tests

1. Crawford, C.C; Raynaldo, D.A.
"Some Experimental Comparisons of True-False Tests and Traditional Examinations,"
School Rebiew 33: 698-706 Nov.1925

were tried out were not familiar with the true-false form. Furthermore, in all the comparisons except one the traditional examinations preceded the others, and in the one in which the true-false test was, by accident, given first, it showed itself distinctly more reliable than the following discussion examination. After mentioning several other factors, their conclusion is that the data they present are not sufficient to warrant a general statement as to which type of test is superior. However, there seems to be little reasonable doubt that, if new-type tests and discussion examinations are constructed with the same degree of care and expertness, and, if the pupils spend the same amount of time working on each, the results on the former will be decidedly more reliable than those on the latter. It is probable that, if several varieties of the new-type examination are combined into one test, the resulting reliability of the total scores will be even greater than if only one form is used. In many of the investigations the comparisons have been made on the basis of a single form only, or, if on several forms, the figures have been reported separately for each.

Pupils feel that many of the ratings they receive on traditional examinations are really unreliable

and unjust. On the other hand, the comparatively high objectivity of new-type tests produces a distinctly favorable re-action. It renders the pupils much more satisfied with the marks which they receive, and enables teachers to justify marks to pupils and their parents much more easily. When pupils compare papers with one another they see that the same response has been scored in the same way, no matter on whose paper it occurred, and thus their confidence in the meaning of marks and the reliability of those they receive is increased. Furthermore, the quality of near-objectivity renders it possible for pupils to score their own answers or those of one another on many occasions, and thus save the teacher considerable work.

Not only because of objectivity in scoring, but for other reasons also, pupils tend to prefer new-type tests. It is true that much depends upon the attitude of the teacher and how the tests are presented to the pupils, but, if the teacher is not prejudiced against the new-type, pupils will almost always favor them. Kinder[1] reports, for example, that of more than two hundred students, all but seven preferred the new-type. Brinkley[2] found that pupils preferred a mixture of essay-type and new-type to all of either one alone.

Other reasons why new-type examinations are

---

1. Kinder, J.S. -"Supplementing our Examinations". Education, 45: 557-566 May, 1925
2. Brinkley, S.G. "Values of New-Type Examinations in the High School with special reference to History" Teachers College, Columbia University, Contributions to Education, No. 161 p.59 1924.

preferred are: that results can usually be known soon after the tests are taken; that less nervousness and fear are aroused; that there is little danger of answers being midunderstood; that the personal likes or dislikes of teachers have practically no opportunity to affect scores, and that there is no visual or writing strain.

A fact more or less dependent upon reliability, and yet distinct from it, is that new-type tests possess greater validity than do discussion examinations. This is caused by their greater objectivity and reliability, and also by the fact that the pupils' answers are very little affected by such factors as their ability in English, handwriting, and so forth. That is to say, the answers are indicative of their knowledge of the subject-matter covered, and not of extraneous abilities which may enter into their answers on essay examinations.

Many of the same writers who have dealt with the question of reliability have also submitted data regarding validity, as also have others. McAfee[1] found correlations of .75 and .79 for new-type tests with a composite measure composed of both new and traditional examination marks, standardized test scores, and teachers' marks. The correlation of discussion examination marks with the same composite was only .66.

1.McAfee, L. O.,- "The Reliability of Non-Standardized Point Tests," Elementary School Journal, 24;579-585 April 1924

It is true in the case of validity as in that of reliability, that not all those who have studied the question are in entire agreement. One of the most careful investigations reported is that of Brinkley[1] who reached the conclusion that with tests of equal length, as measured by time spent in testing, and prepared by teachers with training in the matter of test construction, one type of test yielded practically as good results as another for measuring senior high school achievement in history. With one or two exceptions, he found this true whether the achievement measured was general achievement for the course, ability to think with the materials of the course, or information. He also states that, for measuring general achievement in history, essay examinations are more valid than new-type tests prepared by ordinary high school teachers, and even slightly more valid than those prepared by teachers trained in the construction of the new-type. In the case of new-type tests prepared by Brinkley himself the validity equalled that of essay examinations. For measuring ability to think he found that the two types possessed about the same validity, and for measuring stock of information that the new-type was slightly more valid.

---

1. Brinkley, S.G. "Values of New-Type Examinations in the High School with Special Reference to History." Teachers College, Columbia University p.58

Odell[1] writes the following: "It seems to me that no general conclusion can be drawn as to which type of examination is more valid. The purpose which an examination is intended to serve must be taken into account. For the measurement of stock of information and knowledge of facts, the evidence seems to support the statement that the new examination is more valid than the discussion type. For the measurement of other outcomes of instruction, the data available at present do not warrant the statement that the new-type examination is known to be superior to the traditional type. In other words, each has its particular place and its special functions where it should be preferred to the other."

Since scores upon objective or near-objective tests are helped very little by knowledge that has something to do with the point at issue but does not specifically include it, their use tends to lead pupils to acquire relatively exact and detailed knowledge. Thornton[1] says, "It is evident that pure knowledge must be measured apart from the other factors of historical ability. Only one ability can be tested at a time. Since all teachers will desire to measure information, as well as the other factors, a possible use for informational tests at once appears. In history the specific information must be called for in

1 Odell, C.W. op.cit., p. 196

1 Thornton, E.W.-"The Use of Informational Tests in American History Teaching." 20: 12-16, Jan.1929

the test, if we are to know whether the pupil has it or not."

Also, new-type examinations point out very definitely the particular things which are not known, and thus pave the way for very definite and purposeful remedial instruction. It is felt that objective tests have made a real and valuable contribution toward improvement of instruction, as the teacher studies the test paper as a physician does the findings of his thermometer, and stethoscope;then, having diagnosed, proceeds to modify instruction to remedy the weak places revealed.An essay examination may reveal a general lack of knowledge on a certain topic, but it rarely points out the exact points which need attention.

Pre-Tests in the objective form are now given to history classes by many teachers.

One high school teacher[1] says, "the great value of the pre-test is that it acquaints the teacher with at least, a part of the attitudinal background, or the informational background which each student brought to the subject. Thus the intelligent use of this knowledge might very well determine which groups of facts or ideas would need to be stressed and which not. A true-false pre-test can well be used for this purpose."

---

1. Everett, S.-"Objective Tests the Best Discoverer of Pupil Attitudes," The Historical Outlook, 20:335-337 Nov.1929

Kepner[1] writes that the primary purpose of this kind of test is diadnostic. It aims to measure objectively the extent of the background which a pupil or group of pupils bring to their subject in the secondary school.

New-type tests not only aid the teacher in diagnostic and remedial work, but make it easier for pupils to check up on the results of their own study. It is not bery difficult for a pupil to determine whether he knows certain facts definitely or not, and if he finds that he is ignorant of some of them, to devote further study to those not known.Therefore, new-type tests provide better motivation for study than do discussion examinations.

Workbooks in history have come to the forefront greatly in the past two or three years. Wesley[2] divides them into two classes, the general and the specific.

A workbook of the general type is organized on a topical or chronological basis, and usually contains citations to various texts. So far, this type seems to have been rather restricted in number.

A workbook of the specific type is based upon a published text and parallels it in organization. The specific workbook usually contains topics, problems, projects, exercises, maps, drills, and tests

1. Kepner, T.- "An Aspect of History Testing," The Historical Outlookm15:414-417,Dec.1924
2.Wesley,E.B.-"Workbooks in the Social Studies," The Historical Outlook,22: 151-154,Apl 1931

which are based upon material foundcin the textbook. Some of the specific workbooks contain study-guide sections covering the reading matter in the text. The specific workbook is intended to facilitate the mastery of a specified text.

The idea back of the workbooks is to make the material of the social studies more concrete, of parelleing so far as possible the scientific method of approach.

Workbooks do tend to make the content of the social studies more definite. The student knows what to hunt for and the teacher knows what to ask when the class assembles.

The wrokbooks are designed to absorb at least some of the functions of textbooks, mapbooks, notebooks, scrapbooks, reading books, and written reports.

It would appear that the utility of workbooks increases up to and including the second year in high school, and that it probably declines after the third year in high school.

It is probable that the general workbook can be adapted to more advanced students than the specific workbook, for the former can be less mechanical and can allow greater leaway in procedure.

The tests which are included in workbooks cover

the chapters, and are probably superior to those usually made by the teacher. They are not standardized, and afford nothing more than an objective basis for marking a part of the work, but that is a decided merit.

One point in which new-type examinations possess considerable advantage over traditional ones is in the ease of scoring. Careful and accurate scoring of the answers to traditional questions is relatively difficult amd requires considerable time. By the preparation and use of a list of correct answers which can usually be put in such form that they can be matched with the pupils' responses, the scoring of new-type tests is rendered easy. Not only is time saved, but the type of mental activity engaged in while scoring is much less arduous and tiring than is true in the case of essay examinations.

If it is desired amd practicable, some clerk or other person who does not possess any particular knowledge of the subject-matter covered can score most new-type examinations satisfactorily.

It is also frequently possible to have satisfactory scoring done by the pupils themselves, who may mark their own papers or those of others. They can easily see just what their errors are and also learn how to correct these errors. In most cases it is not

necessary for the teacher to give very much help in this respect, if the pupils are properly motivated so that they have formed the habit of studying their retruned test papers and trying to profit as much as possible thereby. They will ordinarily gain much more benefit from unaided or only slightly aided study of new-type test papers than from that of traditional examination papers. It is impossible for pupils to deceive themselves as to the correctness and quality of their answers.

The fact that the construction of a fairly large number of new-type exercises or items calls for the expenditure of more thought than that of a few essay questions frequently causes teachers to be more careful and thoughtful in so doing. This, of course,leads to the result that more time is required to construct an objective or near-objective test than for an essay examination consuming the same time.

For a very small class this extra amount of time will ordinarily more than offeset that gained in scoring, but for a class of twenty-five, thirty, or forty pupils, this will rarely occur, and it is likely that the time saved in scoring new-type tests will either balance or more than balance the extra amount required in their construction. Even if the total amount of time required is the same, this should be considered a merit of new-

type examinations because a greater proportion of it is spent on construction and less on scoring. In other words, a teacher spends more time in giving consideration to her general objectives, methods, and so forth, and less in what is largely drudgery and mere clerical work. Therefore, the quality of examinations should be improved because of this fact.

It is possible to prepare two or more new-type tests over the same subject-matter which are very nearly equivalent in difficulty, a thing which is practically impossible with essay examinations. If it is desired to give a new-type test of, say, forty items, and to have two forms of the test, eighty items should be prepared. These should then be divided by some random or chance method into two lists of forty items each. The two tests will, in most cases, not be of exactly the same difficulty, but it will be unusual if the difference in difficulty between them is more than a very few per cent.

Moreover, in many cases new-type examinations can be used over again with comparitively slight modifications. Since the number of items contained is comparitively large, pupils cannot expect to make high scores by studying a very small portion of subject-matter, as would be the case if only a few discussion

questions were to be repeated.

Mitchell[1] is of the opinion that "early in the term practice tests should be provided to accustom students to the new-type questions. The first may be a test dealing with ability to follow directions. It should be brief, not more than ten items, and can be corrected at once by the pupils. The second test may well be one of reading comprehension."

3) Dis-advantages of new-type examinations-----

One of the dis-advantages of the new-type examination is that it is harder to make and to give than the traditional type. Since it is composed of more questions or exercises the teacher takes more time to prepare it. It is probably true, also, that the degree of mental effort required in its construction is more, although this may not hold if the persone making it is equally familiar with the new-type test, and if the traditional examination constructed is of as high a degree of merit as the other.

Moreover, in most cases new-type tests generally require that a copy be placed in the hands of each pupil in order to be effective. This requirement is frequently a very practical hindrance to their use, since it is sometimes absolutely impossible, and frequently decidedly difficult, for a teacher to provide the necessary number of copies. Many schools, especially small ones,

1. Mitchell, Elene. Teaching Values in New-Type History Tests, p.95

do not have any sort of device, such as a mimeograph, or hectograph, by which a number of copies can be made. Even in the case of schools which do have such devices, it is not always easy to secure the desired number of copies. For a very small class it may be practicable to use carbon-copies made upon the typewriter, but for classes of ordinary size, that is hardly practicable, requiring too great an amount of labor.

Although much has been said and written concerning the new-type tests, they have been scarecely heard of by many teachers, and are not understood as to purposes, limitations, and administration, by many others.

It seems probable that those who favor the traditional examination are correct in their assertion that it tests reasoning and most other thought processes except memory better than do new-type tests. The latter tend to measure only knowledge of facts acquired, and that often in rather disconnected fashion. Such qualities and mental activities as originality, initiative, organization, interpretation, anaysis, discrimination, judgment, subtlety, and so forth, are, it is said, only slightly if at all measured by new-type examinations.

Krey[1] writes, "thus far, we have not been able to gain any help from the objective tests in arriving at opinions of "subjective qualities" of the students."

In most varieties of the new-type examinations pupils are in some form or other given a number of possible answers from which to select the correct ones. In other words, they are not thrown upon their own resources to the same extent as by discussion questions. Knowledge which is only marginal or hazy is frequently sufficiently quickened by the suggested answers that the correct responses are recognized, whereas there is usually no result of this sort in connection with the essay examination. The argument may be made, however, that it is not altogether undesirable that this be the case, that is, that some tests secure more or less suggested responses. On the whole it appears that new-type tests are inferior to discussion examinations on this point, but not so far inferior as has sometimes been asserted.

A strong objection commonly made to the new examination is that it encourages guessing. This is especially charged against alternative tests, in which pupils know that they have one chance out of two of guessing right in each particular case, but also to some extent against multiple-answer tests, matching tests, and several other varieties. It is undoubtedly

---

1. Krey, A.C.- "What Does the New-Type Examination Measure in Histroy," The Historical Outlook, 19: 159-162 April, 1928

true that it is easier for pupils to give brief responses than to write rather long discussions, if they know little or nothing about the matter at issue in either case. However, if the tests are properly administered, including satisfactory directions for the pupils, and also properly scored, it is not apparent that the amount of guessing which occurs is great enough to be a very serious fault. One item in satisfactory directions should be a statement strongly advising or directing pupils not to guess, that is, not to record an answer unless they are at least fairly sure it is correct. However, if one wishes to obviate the possibility of pupils profiting by guessing in spite of the ordinary scoring methods to be described later, he may well do something of the sort suggected by Christensen[1]. This is, that after one type of test, such, for example, as a true-false one, has been given, it be followed fairly soon, that is within a day or two, by one of another type, perhaps multiple-answer, covering the same material and even corresponding item for item with the first test. The two tests should then be scored together and credit given only for those items correctly answered in both. Even apart from its tendency to reduce guessing, such a repetition is occasionally desirable.

---

1. Christensen, A.M.-"A Suggestion as to correcting Guessing in Examinations." Journal of Educational Research 14: 370-374 Dec.1926

An objection frequently urged against new-type tests is that they, or at least several of the most common varieties of them, tend to confuse the pupil as to what he really knows, or even to teach him erroneous facts. This charge is especially made against true-false exercises, since practically half of the statements contained therein are false. It is also made against the multiple-answer type, in which several of the suggested answers are incorrect; against the matching type, in which the pupil may form wrong combinations which will tend to remain in his memory; against the incorrect statement type for the same reason as in the case of the true-false type, and so forth. In reply this criticism there are at least three points to be made.

In the first place, material which the pupils have not already studied and supposedly learned in the correct form should not be covered or presented in tests. If something has already been well learned, this knowledge will not be disturbed or confused by seeing a false statement concerning the matter.

In the second place, it may be true that some confusion regarding facts only partially mastered may be caused, but this should be satisfactorily taken care of and corrected by the teacher in her discussion of the errors made upon the test.

Finally, it should be recognized that in life outside the school individuals are very frequently called upon to distinguish true statements from false ones, valid arguments from invalid ones, to select the best of several possibilities, or to do something else resembling very closely some variety or other of new-type tests. It is therefore highly desirable that some training along these lines be given pupils in school, and it is eminently worthwhile to risk the confusion of ideas and knowledge, which may occur to a limited degree, in the endeavor to avoid much more serious confusion later and to develop critical ability in the ordinary affairs of life.

The statement has been made that objective or near-objective tests are too artificial in that they do not resemble life's situations or problems in one important particular. This is that the problems met in life outside the school are such that there is rarely one and only one correct solution and all others wrong, but instead there are frequently several solutions of approximately equal merit, or perhaps several of which one is slightly better than another, the second slightly better than a third, and so on. It is therefore argued that pupils should not become accustomed to looking for answers

or solutions which are absolutely right or wrong, but should be trained as much as possible in dealing with situations in which all or practically all of the elements or factors are subjective. Although there is some truth in the contention just stated, it is not apparent that traditional examinations as ordinarily administered are of much if any more value in giving training of the kind desired than are new-type tests. Probably, if traditional examinations were administered with this end in view, they could be made to yield considerably greater returns along this line than is true at present, and also greater ones than would come from objective tests. On the other hand, some varieties of the latter, such as the multiple-answer type with several answers of varying degrees of merit, do give training of the type specified above.

4) Five most generally useful varieties of the New-Type examination forms------

1. Single-answer test.

Directions----The correct answer to each question below is a single word. If you know, or think you know, the answer toa question, write it upon the blank line in front of that question. Do not write more than one word upon any line.

_____1. What other term was applied to the "Carpet-baggers" of the North?

_____2. Who was the author of the Declaration of Independence?

2. Multiple-answer test.

Directions----Each of the questions below is followed by five suggested answers of which one is right. If you think you know which one is right, place the letter before it on the short blank line in front of that question.

_____1. The most important modern state to adopt a policy of free trade was_____
A.The United States. B.Russia.
C.Italy. D.England. E.France

_____2. Of the following men, the one who is famous as a political pholosopher is_____
A.John Law. B.Eli "hitney. C.Richard Arkwright. D.John Locke.
E.Sir Walter Raleight.

3. Alternative Test.

Direction-----Below are a number of statements of which about half are true and the other half false. In the case of each statement that you think is true, place a plus (+) mark on the blank line in front of it, and in the case of each that you think is false, place a minus(-)sign. To show just how this is to be done, the first two sentences have been marked. The first one is true, as it has a plus sign in front of it, and the second is false, so it has a minus sign before it. If you do not think you know whether a statement is true or false, do not guess, but omit it and go on to the next one.

_____ 1. Under the Articles of Confederation, Congress had no power to collect taxes directly from individuals.

_____ 2. The acquisition of Louisiana more than tripled the territory of the United States.

_____ 3. President Wilson promised the Filipinos that the United States would grant them complete independence before 1930.

4. Completion Test.

Directions----- Each of the blanks in the paragraph below represents the omission of one word. If you think you know the word that should ne there, write it on the blank. Do not in any case write more than a single word on one blank.

Our Federal Government has three branches, the_____ the_____, and the_____.The President is at the head of one branch, Congress at that of another, and the_____ at that of the third. The President is assisted by his_____ in which there are_____ members. Congress consists of _____Senators from each State and Representatives whose number is determined by the_____ of the various States.

5. Matching Test.

Directions---- Below is a group of items, divided into two columns. Each item in the left-hand column is numbered. Each item in the right-hand column is followed by parentheses. Place in the parentheses the number of that item in the left-hand column that matches with the item in the right-hand column.

1. Spanish-American War	Brigham Young ( )
2. Organized labor.	Elias Howe ( )
3. Sewing Machine.	James J.Hill ( )
4. Mormons.	Alexander H. Stephens ( )
5. Civil Service Reform.	Samuel Gompers( )
6. Railroads.	
7. Southern Confederacy.	

Mitchell[1] recommends that "especially in the social studies it is important to adapt the form of question to the material used, and to phrase questions to emphasize associations with other facts instead of with mere words."

1. Mitchell, Elene, Teaching Values in New-Type History Tests. p. 63

5) Conclusion---------

New-type tests are, as is shown by a considerable mass of evidence, more reliable than traditional examinations, both because they secure better samplings of pupils' ability and knowledge, and because their scoring is relatively objective.

Among the other advantages which they possess are that pupils usually prefer them, and are better satisfied with the marks which they receive, emphasis is placed upon exact and accurate knowledge, knowledge of the subject being tested is measured without being mixed with ability in language, handwriting, and so forth, speed can be measured when desired, scoring is easier, more thought is usually required in their construction, bluffing is more difficult, and two or more forms of practically equivalent difficulty can be prepared.

Pre-tests in the new-type form are being used for diagnostic purposes, and the results from these tests are used for the improvement of instruction through diagnosing the shortcomings and difficulties of pupils.

Practice tests are being urged to train the pupil beforehand in the use of new-type tests.

Workbooks with objective tests therein have come greatly to the forefront.

The five most generally useful varieties of new-type test forms are: single-answer; multiple-answer; alternative; completion, and matching.

Chapter 8

Summary

Chapter 8

Summary------

The value of educational measurement in general is no longer a matter of doubt. Immeasureable benefits to all pupils and persons related to the schools, including the community at large, have come from the results of testing.

The past quarter of a century has witnessed the rise of educational measurement to the plane of conscious striving for objective, impartial and comparative means for portraying the absolute and relative achievement of pupils. The measurement of achievement has been admittedly the principal reson for examinations. All seem to be agreed that the first purpose of a test or examination is that of ascertaining the degree to which individual pupils have profited by instruction. That purpose should be attained by accurate, objective and impartial measurements. It should ever be kept in mind that it is the pupil who is being tested or measured, therefore, such factors as subjective judgments should be reduced to a minimum, if the results are to be fair to all.

Scientific measurement has passed the experimental stage. Because of its importance all teachers

should have some knowledge of this subject. To be thoroughly equipped for their positions they must know something of the aims, methods, materials, results, and general educational implications of this important work.

The four principal and most important types of measurement in history in the modern high school are: oral questioning, the traditional examination, the standard test, and the objective or new-type examination. With so little of a final character known about the relative merits of the above methods, the only course open to teachers and educators is to consider both the logic and the growing body of experimental findings supporting or undermining the value of each.

Although, as has been stated, the first purpose of a test is that of measuring the achievement of pupils, history tests in the high school are also being used for the purpose of promotion, to test the efficiency of a teacher, to examine pupils for purposes of locating beginning points in teaching, for determining school status, they may promote review and recall, they may be used for placement of pupils or for classification, they may be given in order to diagnose difficulties of pupils, and they may in that way provide a basis for remedial measures. They

may be used for the motivation of school work, and for the promotion of real and not artificial interest. History tests may provide pupils with objective goals for school study. They may measure memory. Most important of all, to the extent to which their influence may be felt, history tests may improve teaching.

For many workers in educational measurements it is more worthwhile to be able to recognize a good test by its earmarks than to know intimately the many tests,- good, mediocre, and poor, - that exist to-day. The chief criteria which should be kept in mind in selecting history tests in the high school are: validity, reliability, objectivity, norms, duplicate forms, scaling, ease of administration, and cost. The two fundamental criteria for selecting a history test are validity and reliability,-all others are subordinate. Objectivity in scoring is one important factor in reliability. Validity refers to whether or not a test accomplishes its purpose. A test is valid when it measures the ability or characteristic that it is supposed to measure. A test is reliable when it measures whatever it does measure accurately; or, in other words, if the same results are secured when it is given two or more times to the same pupils. Among the chief factors that affect reliability are objectivity, the length of a test, even-

ness of scaling of the test elements, and the directions for giving and scoring. A test or score is objective, if it is not influenced by the personal opinion or judgment of the person doing the scoring-that is, if all competent scorers agree. For actual school practice lerely placing the items in an order of difficulty suffices in a comprehensive objective tests. A test should go to low and to high enough limits. Though not of great theoretical importance, the ease of administering a test and its cost are of practical importance.

As for the value of oral questioning for measurement in history in the high school, we must conclude that oral questioning is more logically a part of initial instruction than of final measurement. Strictly speaking, oral questioning does not usually constitute an examination. Oral examinations are sometimes employed, but their value for the more serious and final determination of achievement is doubted. Oral questioning is primarily instructional; its value for measurement is more subordinate.

Although there has been much recent unfavorable criticism of traditional examinations, they should not be entirely discarded in favor of new-type tests, but should be used on some occasions. Each of the

two general types just mentioned has its peculiar merits and advantages, and should be employed when it best fulfills the desired end. Traditional examinations are usually much easier to prepare, test a number of mental processes better than does the new-type, do not offer a great opportunity for guessing, and perhaps not for cheating, are not as liable to the danger of confusing the pupil, and in several minor ways are to be preferred. On the other hand, traditional examinations are, as is shown by a considerable mass of evidence, less reliable than the new-type tests because they do not secure as good samplings of pupils' abaility and knowledge, and because their scoring is relatively subjective and more difficult. Bluffing is also less difficult. The "traditional" examination, also known as the "essay" or "discussion" examination, has been almost exclusively employed in the high school until the last few years, and still perhaps more usual than any other.

Standardized examinations have just completed the first quarter century of their existence. The curve of their use is rising rapidly, and this use has come to be an integral part of the work of teachers in many systems, both large and small. In general usage, the adjective "standardized" or "standard" implies

that the test in question has been carefully constructed according to certain general principles, and embodies exercises of such forms that pupils' responses are relatively, if not absolutely, objective. It also implies that the test has been widely enough given that the results therefrom indicated what might be expected of pupils of a given age, grade, or other homogeneous group.

It was as a relief from the admitted weakness of the traditional examination,- that of subjectivity of marking,- that the standard test was introduced. Other principal aims are: to control or standardize the conditions of the examination period, to provide norms or standards for evaluating and interpretating the scores of individual pupils. The last aim is the most difficult, and, on the whole, decidedly the least important.

Standardized tests have a legitimate place in high school measurement in history, and should not be wholly displaced by other types of measurement. Certain advantages accrue from the standardized tests which the others cannot supply. They are characterized by the care exercised in their making, by the objectivity of their scoring, by the comprehensiveness within the field measured, by the universal character

of their content, and by the fact that they are accompanied by norms, or standards of achievement. Standardized tests are generally more objective, reliable, and valid than either the traditional type or the new-type examination. Also, their use saves time in both preparation and scoring. Equivalent forms provided by most good standardized tests insure greater accuracy in measuring progress. But the standardized test should not be the sole element in a measurement program. It has its limitations as well as its advantages. Among the former is the serious one that standardized tests are not sufficiently flexible to satisfactorily meet local needs. They are non-adaptable to local school curricula.

The new examinations or, better, the new-type test, is the name commonly given to tests or exercises, generally constructed by a teacher for her own use, that make use of the forms and scoring methods of standardized tests so as to possess relatively high objectivity, but have not gone through a process of careful trying out of material included. New-type tests are more reliable than traditional examinations, both because they secure better samplings of pupils' ability and knowledge, and because their scoring is relatively objective. Among the other advantages which

they possess are that pupils usually prefer them, and are better satisfied with the marks which they receive, emphasis is placed upon exact and accurate knowledge, knowledge of the subject being tested is measured without being mixed with ability in language, handwriting, and so forth, speed can be measured when desired, scoring is easier, more thought is usually required in their construction, bluffing is more difficult, and two or more forms of practically equivalent difficulty can be prepared.

Pre-tests in the new-type form are being used for diagnostic purposes, and the results from these tests are used for the improvement of instruction through diagnosing the shortcomings and difficulties of pupils.

Practice tests are being urged to train the pupil beforehand in the use of new-type tests.

Workbooks with objective tests therein have come greatly to the forefront.

The five most generally useful varieties of new-type tests are: single-answer, multiple-answer, alternative, completion, and matching.

The teaching of history through testing, in the high school should utilize a complete testing program

Such a program should generally consist of the traditional examination; standardized tests, and the so-called new-type examination. Each has its peculiar merits and advantages, and should be employed when it best fulfills the desired end.

Appendix.

How To Secure Tests and Directions For Their Use.

Appendix

How To Secure Tests and Directions For Their Use.

Tests are changing at a phenomenal rate and changing for the better. It is the function of frequent bulletins issued by Book Companies and bureaus of research to inform educators of the latest and best tests. All the important centers which distribute testing material are prepared to send free or practically free literature describing their tests. More than this, they are glad to give expert advice as to the test or tests which it is best to use in a particular situation. Finally, they are prepared, for a small charge, to send for inspection sample tests. Again, the bureaus which issue tests usually fo, and always should, send with the tests which have been ordered, a leaflet giving detailed directions for applying and scoring the tests, for tabulating results, and for computing pupil and class scores. The directions usually include norms for the test and, frequently, suggestions for the uses of results. As a precaution, the individual, when writing for tests, should request that all necessary directions for properly using them be sent.

In the field of history a number of tests are now available for use in the high school. It would be imposible to describe in detail all available

history tests. The following is a list of the names of some available standardized history tests for use in the high school, together with the names of their publishers.

Barr's Diagnostic Tests in American History.
Public School Publishing Company (1918)

Columbia Research Bureau American History Test
World Book Company (1926)

The Gregory American History Tests
C. A. Gregory, University of Cincinatti (1923)

Harlan Test for Information in American History
Public School Publishing Co pany (1917)

Iowa General Information Test
Bureau of Educational Research and Service, University of Iowa (1927)

Pressey-Richards Test for the Understanding of American History
Public School Publishing Company (1922)

The Gregory-Owens Tests in Mediaeval and Modern History
C. A. Gregory, Univ. of Cincinatti (1926)

Pressey Tests of Historical Judgment
Public School Publishing Company (1924)

Van Wagenen Reading Scales
Public School Publishing Company (1922)

American Council European History Test
World Book Company.

The World Book Company, Yonkers-on-Hudson, New York, and the Public School Publishing Company, Bloomington, Illinois, are perhaps the largest publishers of standardized history tests.

The World Book Company has just issued a book-

let entitled "Bibliography of Tests for Use in Schools", which sells for ten cents. This booklet gives tests sold by other agencies than themselves.

A Bibliography of Books and Sources Used, in part, in the Organization of this Thesis.

1) Brinkley, Sterling G.
"Values of New-Type Examinations in the High School with Special Reference to History"
Teachers College, Columbia University Contributions to Education, No.161

2) Christensen, A.M.
"A Suggestion as to Correcting Guessing on Examinations,"
Journal of Educational Research
14: 370-374, Dec. 1926.

3) Colvin, Stephen S.
An Introduction to High School Teaching (1922
The Macmillan Company, New York.

4) Crawford, C.C., and Raynaldo, D.A.
"Some Experimental Comparisons of True-False Tests and Traditional Examinations,"
School Review
33: 698-706, Nov. 1925.

5) Douglass, H.R.
Modern Methods in High School Teaching (1926)
Houghton Mifflin Company, Boston.

6) Elston, Bertha
"Improving the Teaching of History in the High School through the use of Tests,"
The Historical Outlook
14: 300 - 305, Nov. 1923

7) Everett, Samuel
"Objective Tests the Best Discoverer of Pupil Attitudes,"
The Historical Outlook
20: 335-337, Nov. 1929

8) Hardy, Ruth E.
"New Types of Tests in Social Science,"
The Historical Outlook
14: 326-328, Nov. 1923

9) Hill, H. C.
"The Use of Tests in the Teaching of the Social Studies,"
The Historical Outlook
20: 7 - 10, Jan. 1929

10)Kelley, T. L.
"Note on the Reliability of a Test: a Reply to Dr.Criticism,"
Journal of Educational Psychology
15: 193-204, April, 1924

11)Kepner, Tyler
"An Aspect of History Testing,"
The Historical Outlook
15: 414-417, Dec. 1924

12) Kepner, P. T.
"A Survey of the Test Movement in History,"
Journal of Educational Research
7: 309-325, April, 1923

13) Kinder, J. S.
"Supplementing our Examinations,"
Education
45: 557-566, May, 1925

14) Krey, A. C.
"What Does the New-Type Examination Measure in History?"
The Historical Outlook
19: 159-162, April, 1928

15) Lincoln, Edward A.
Beginnings in Educational Measurement (1924)
J.B.Lippincott Company, Chicago, Ill.

16) Lindquist, E.F.
"Factors Determining Reliability of Test Norms,"
Journal of Psychology,
21: 512-520, Oct. 1930

17) Lindquist, E.F., and Anderson, H.R.
"Objective Testing in World History,"
The Historical Outlook
21: 115-122, March, 1930

18) McAfee, L .O.
"The Reliability of Non-Standardized Point Tests,"
Elementary School Journal
24: 579-585, April, 1924

19) Mc Call, William A.

How to Measure in Education (1922)

The Macmillan Company, New York.

20) Mitchell, Elene

Teaching Values in New-Type History Tests (1930)
World Book Company, Yonkers-on-Hudson, New York

21) Monroe, W. S.

"Written Examinations and Their Improvement,"

The Historical Outlook

14: 211-219 June 1923
14: 306-318 Novr. 1923

22) Moyer, F. E.

"New Types of History Tests,"

The Historical Outlook

14: 323-324 Novr. 1923

23) Odell, Charles W.

Educational Measurement in High School (1930)

The Century, Company, New York.

24) Odell, Charles W.

Traditional Examinations and New-Type Tests (1928)
The Century Company, New York

25) Parker, Samuel C.

Methods of Teaching in High School (Revised Edition 1920)
Ginn and Company, Boston, Mass.

26) Pressey, Sidney L. and Pressey, Luella Cole

Introduction to the Use of Standard Tests (1923)
World Book Company, Yonkers-on-Hudson New York

27) Ruch, G. M.

The Objective or New-Type Examination (1929)
Scott,Foresman and Company, New York

28) Ruch, G. M.

The Improvement of the Written Examination (1924)
Scott,Foresman and Company, New York

29) Ruch, G.M. and Stoddard, George D.

Tests and Measurements in High School Instruction (1927)
World Book Company, Yonkers-on-Hudson New York

30) Rugg, E. M.

"Character and Value of Standardized Tests in History,"
School Review,
27: 757-771, Dec. 1919

31) Russell,Charles

Classroom Tests (1926)
Ginn and Company, Boston, Mass.

32) Russell,Charles

Standard Tests (1930)
Ginn and Company, Boston, Mass.

33) Sangren,P. V.

"The Present Status of Measurement in the Social Studies,"
The Historical Outlook
21: 279-283, Oct. 1930

34) Starch, D., and Elliott, E.C.
"Reliability of Grading Work in History,"
School Review
21: 676-681, Dec. 1913

35) Symonds, Percival M.
Measurement in Secondary Education (1927)
The Macmillan Company, New York

36) Symonds, Percival M.
"A Study of Extreme Cases of Unreliability"
Journal of Educational Psychology
15: 99-106, Feb. 1924

37) Smith, H. L., and Wright, W.W.
Tests and Measurements (1928)
Silver, Burdett and Company, Boston, Mass.

38) Thorndike, E.L., Bregman, E.A., and Cable, M.V.
"The Selections of Tasks of Equal Difficulty
by a Consensus of Opinion,"
Journal of Educational Research
9: 133-139, Feb. 1924

39) Thornton, E. W.
"The Use of Informational Tests in
American History Yeaching,"
The Historical Outlook
20: 12-16, Han. 1929

40) Trabue, Marion Rex
Measuring Results in Education (1924)
American Book Company, Boston, Mass.

41) Tryon, R.M.
"Standard and New-Type Tests in the Social Studies,"
The Historical Outlook
18: 172-178, April, 1927

42) Wesley, E. B.
"Workbooks in the Social Studies,"
The Historical Outlook
22: 151-154, April, 1931

43) Wilson, G. M.
"Criteria of a Standardized Test,"
Educational Review
71: 138-141, March, 1926

44) Wood, Ben D.
"Studies in Achievement Tests,"
Journal of Educational Psychology
17: 125-139, Feb. 1926

45) Wood, Ben D.
"The Measurement of College Work,"
Educational Administration and Supervision
7: 301-334, Sept. 1921

BOSTON UNIVERSITY
1 1719 02575 1159